DOCUMENT RESUME

ED 067 402 TM 001 797

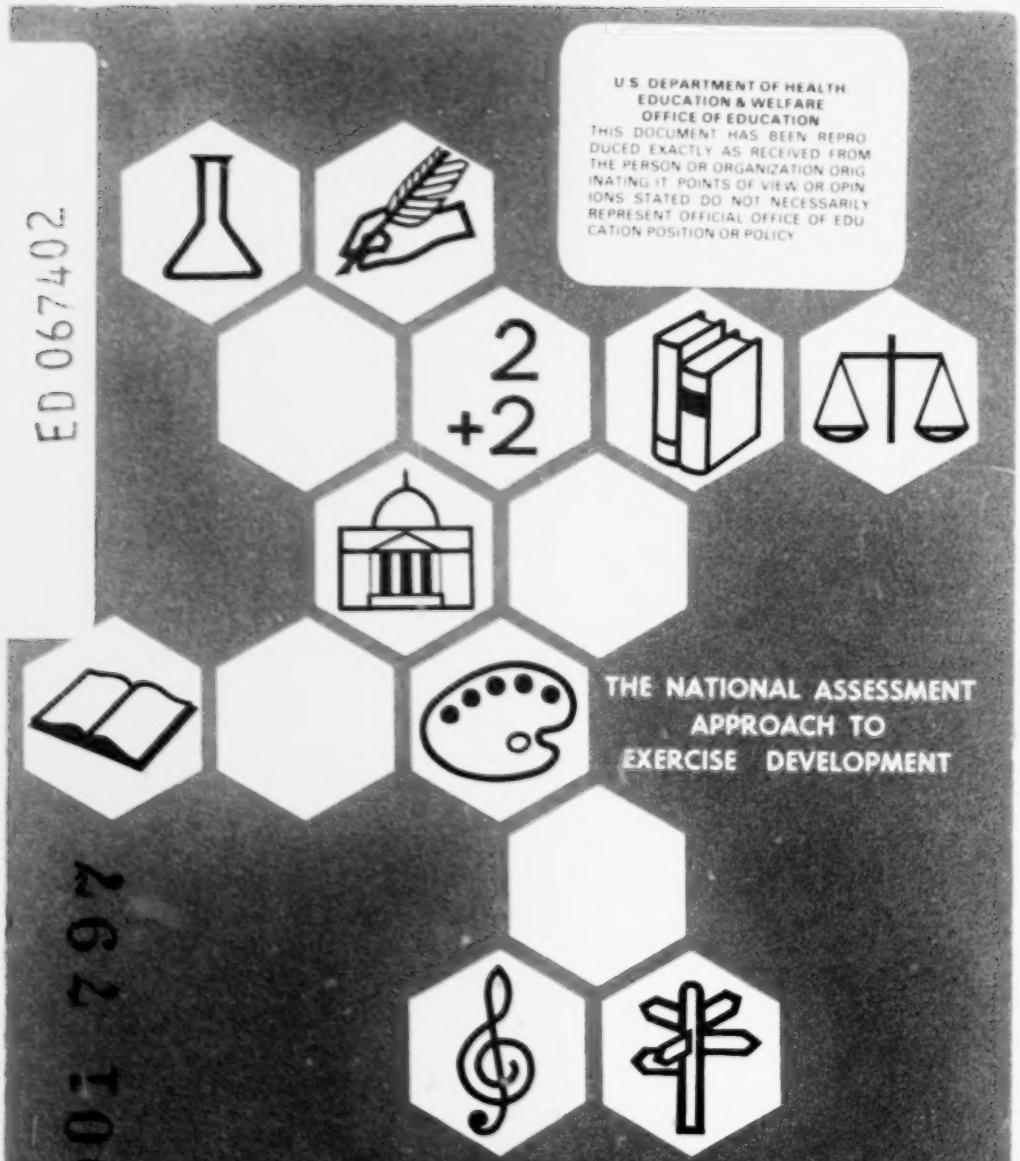
AUTHOR	Finley, Carmen J.; Berdie, Frances S.
TITLE	The National Assessment Approach to Exercise Development.
INSTITUTION	National Assessment of Educational Progress, Ann Arbor, Mich.
SPONS AGENCY	National Center for Educational Research and Development (DHEW/OE), Washington, D.C.
PUB DATE	70
GRANT	OEG-0-9-08771-2468(508)
NOTE	143p.
AVAILABLE FROM	National Assessment Staff Offices, Room 201A Huron Towers, 2222 Fuller Road, Ann Arbor, Mich. 48105 (Single copy $3.00; orders of $10 or more, 20 percent discount)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Academic Achievement; Educational History; Educational Testing; *Evaluation Methods; Group Tests; *Measurement Instruments; *National Competency Tests; *Standardized Tests; *Test Construction
IDENTIFIERS	NAEP; *National Assessment of Educational Progress

ABSTRACT

The history of the development of National Assessment exercises from the project's inception in 1964 to the present is provided in this monograph. The chapter titles are as follows: I. Introduction; II. Rationale and Criteria for Writing Exercises for National Assessment; III. Initial Reviews; IV. Initial Studies; V. Subject Matter Reviews; VI. Other Studies; VII. Final Reviews and Selection; and VIII. New Directions in Exercise Development. A glossary and references are included. (For related documents, see TM 001 793 and TM 001 789.) (DB)

ED 067402

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY



THE NATIONAL ASSESSMENT APPROACH TO EXERCISE DEVELOPMENT

TM 001 797

NAEP

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
A PROJECT OF THE EDUCATION COMMISSION OF THE STATES

ED 067402

THE NATIONAL ASSESSMENT APPROACH TO EXERCISE DEVELOPMENT

Carmen J. Finley
and
Frances S. Berdie

National Assessment of Educational Progress

Ann Arbor Offices:
Room 201A Huron Towers
2222 Fuller Road
Ann Arbor, Michigan 48105

Denver Offices:
822 Lincoln Tower
1860 Lincoln Street
Denver, Colorado 80203

A Project of the
Education Commission of the States

This publication was prepared pursuant to Grant No. OEG-0-9-08771-2468(508) with the National Center for Educational Research and Development, Office of Education, U. S. Department of Health, Education, and Welfare. Grantees undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

National Assessment of Educational Progress, 1970

Library of Congress Catalog Card Number 70-129300

Single copies $3.00
Orders of $10. or more of any National Assessment publications, 20% discount

National Assessment Staff Offices

Room 201A Huron Towers

2222 Fuller Road

Ann Arbor, Michigan 48105

TABLE OF CONTENTS

Foreword i

Note on the Project Title ii

I. Introduction 1
- Historical Overview 1
- Early Plans 3
- The Organization of the Monograph 5
- Example of Exercise 7

II. Rationale and Criteria for Writing Exercises for National Assessment 11
- Objectives: Purpose and Process 11
- The Development of Exercises 14
  - Content Validity 15
  - Clarity 17
  - Type of Exercises and Format 17
    - Scoring Problems 19
    - Format Problems 19
  - Clustering 25
  - Directionality 26
  - Difficulty Level 28
  - Content Sampling and Number of Exercises 30
  - Overlap Between Ages 33
  - Special Concerns for Age 17 33
- Summary 33

III. Initial Reviews 35
- Review of Exercises for Offensiveness 35
- Mail Review of Exercises by Subject Matter Specialists 47

IV. Initial Studies 53
- The 90 Percent Study 55
- Feasibility Studies 60
  - Children in School 61
  - Adults and Out-of-School Youths 64
- Summary 66

V. Subject Matter Reviews 67
First Subject Matter Review Conferences 67
Second Subject Matter Review Conferences 71
Special Problems 73

VI. Other Studies 76
The Mathematics Study 76
The Choices Study 80
Final Field Tryouts Before Assessment 85
Group Administered Exercises 86
Scoring 88
Selected Reactions Offered by Field Coordinators 89
On Content 89
On Tape and Timing 90
On the Use of "I Don't Know" 90
On Format and Motivation 91
Reactions of Students 91
Anecdotes 94
Individually Administered Exercises 94
How Did Tryouts Affect Exercise Development? 97

VII. Final Reviews and Selection 98
Review of Exercises for Meaningfulness 98
Final Review by the Technical Advisory Committee .. 103
USOE Review for Invasion of Privacy 106
The Selection Process 109
Summary 110

VIII. New Directions in Exercise Development 111
Five Phases 112
Objectives and Prototype Exercises 114
Prototype Exercises 116
The Preparation of Exercises 126
Reviews 127
Field Testing 127
Final Reviews and Selection 128
Summary 128

Glossary 130
References 135

FOREWORD

The intent of this monograph is to record the history of the development of National Assessment exercises from the inception of the project in 1964 to the present time. Major issues which relate to objectives or field operations are mentioned only incidentally and where they are important to the development of assessment exercises. Other monographs are planned which will treat these two aspects more thoroughly.

The information contained herein was gathered from the National Assessment files and the memories of persons associated with the project from its beginning. The authors are indebted to Ralph W. Tyler, John W. Tukey, Jack C. Merwin, Arleen S. Barron and Daryl G. Nichols who read the manuscript in its entirety and made many helpful suggestions. The manuscript was also read in its entirety by Reba E. Sones from the point of view of the lay reader. We are grateful to her for her many ideas and editorial suggestions.

In addition, helpful suggestions on specific points were received from Lee J. Cronbach, Lyle B. Jones, Robert P. Abelson, the late Herbert S. Conrad, Burton E. Voss, Jerry L. Walker, Thomas R. Knapp and Jason Millman. We also wish to thank Lynn Levinson and Lisa MacDonald for their editorial assistance, Charlotte Hayes for her research assistance, and Irene Kowalski and Dean Speaks for their patience in typing the many versions through which the manuscript passed.

Last, but not least, we are most apppreciative of the assistance and encouragement received from Staff Director Frank B. Womer and from Eleanor L. Norris, Director of Information Services.

We hope this record will be helpful to others who are interested in the National Assessment model.

Ann Arbor, Michigan
July, 1970

C.J.F.
F.S.B.

A Note on the Project Title

The reader will note some variation in the terminology used in referring to the National Assessment Project and its governing body. The first governing group which was appointed in 1964 was the Exploratory Committee on Assessing the Progress of Education (ECAPE). In the summer of 1968, the term *Exploratory* was dropped from the title as the original planning and developmental phase of the project gave way to actual operation and the Committee (CAPE) expanded its membership. In July 1969, governance was shifted to the Education Commission of the States (ECS) and the official title of the project became National Assessment of Educational Progress (NAEP). The term *National Assessment* is a general one and is used in referring to any phase of the project.

CHAPTER 1

Introduction

In the summer of 1964, with funding by the Carnegie Corporation of New York, the Exploratory Committee on Assessing the Progress of Education (ECAPE) was appointed to:

1. determine how a national assessment of educational progress could be designed,
2. develop and test instruments and procedures for the assessment, and
3. develop a plan for conducting the assessment.

The focal point of this report is the development of instruments which will be used in the National Assessment of Educational Progress (NAEP).

Historical Overview[1]

During Francis Keppel's tenure (1962-65) as U.S. Commissioner of Education, he became concerned that the original charter in 1867 established as one of the duties of the U.S. Office of Education (USOE) the determination of educational progress in the several states. The only information available had been primarily input variables, such as dollars spent, the buildings occupied and the number of teachers employed. For the first time, attention was turned to the problem of finding out how much has been learned and what progress is being made in education in the United States. After a number of conferences and discussions initiated

[1]Only a brief history of the project is presented here. For a more comprehensive report the reader should consult:

Department of Elementary School Principals, NEA. *National assessment of educational progress: some questions and comments.* (Rev. ed.) Washington, D.C.: Author, 1968.

Merwin, Jack C. and Womer, Frank B. Evaluation in assessing the progress of education to provide bases of public understanding and public policy. In *NSSE yearbook. Educational evaluation: new roles, new means.* Chicago, Illinois: University of Chicago Press, 1969.

Womer, Frank B. *What is national assessment?* Ann Arbor, Michigan: National Assessment of Educational Progress, 1970.

by Commissioner Keppel, John W. Gardner, President of the Carnegie Corporation, asked a distinguished group of Americans to form the Exploratory Committee under the chairmanship of Ralph W. Tyler (Director of the Center for Advanced Study in the Behavioral Sciences, Stanford, California) to consider development of an assessment program which would provide benchmarks of educational progress as a basis for evaluating the changing educational needs of our society over the years.

The Exploratory Committee presented a plan designed to answer two questions: (1) What are the current educational attainments of our population? (2) What change is there in the level of attainments over a period of time? While these goals may not seem unusual as compared to other testing programs, the method by which they were to be achieved in National Assessment contained a number of unique aspects.

Four years of work, financed by private foundations, have gone into defining goals and developing measuring instruments to answer these questions. This work has been done in consultation with leading educators and interested laymen. Ten subject matter areas have been selected for initial assessment: Art, Career and Occupational Development,[2] Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies and Writing. Exercises for the 10 subject areas are in various stages of development: three areas (Citizenship, Science and Writing) were assessed in the first year of administration (1969-70); the other seven are scheduled to be assessed in subsequent years. Additional areas probably will be developed in future years.

Exercises will measure knowledges, skills and attitudes of groups, not individuals. Results in each subject will be reported by:

Four age levels [9, 13, 17 and young adult (26-35)]
Four types of communities (large city, urban fringe, middle-size city and rural-small town)
Four geographical regions (Northeast, Southeast, Central and West)
At least two socio-educational levels
Race (Black, other)
Sex

Dr. Tyler, in writing about the purposes of National Assessment, states:

> The National Assessment is designed to furnish information to all those interested in American education regarding the educational

[2]Formerly called Vocational Education

achievements of our children, youth and young adults, indicating both the progress we are making and the problems we face. This kind of information is necessary if intelligent decisions are to be made regarding the allocation of resources for educational purposes.[3]

For the first time in American education there is a plan to systematically sample the skills, knowledges and attitudes of youth and to report the results to all people involved directly or indirectly in the ongoing process of improving education.

Early Plans

At first, the preparatory work for the assessment was seen as a process requiring about two years of work. At one of the early planning sessions, a time schedule was proposed which gave June 1966, as the date when the preliminary work would be completed and the materials could be turned over to a national commission which would do the actual assessment.

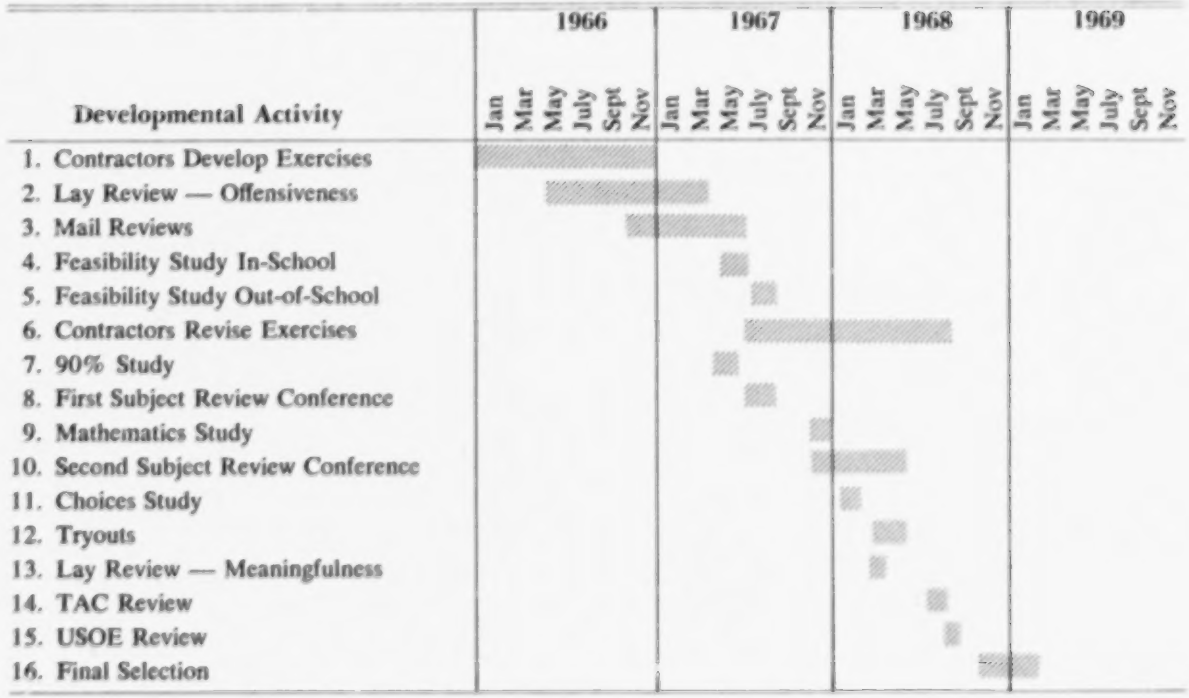
The original plans called for contracting with private measurement agencies to work with teachers and scholars in the development of educational objectives in each of 10 subject areas. These objectives were then to be reviewed by informed lay people representing different geographic areas and types of communities, in order to be sure the objectives were ones considered important by lay people. After the establishment of objectives, the contractors were to develop exercises[4] which would sample the objectives. The exercises were to receive two reviews: (1) by lay people to be sure that no exercise would be included which was potentially offensive to any large group of people and (2) by mail by experts in each of the subject areas to assure that each exercise did assess the educational objective for which it was written, and that it had no identifiable flaws. After the two reviews were completed and needed revisions were made, exercises with different formats were to be tried out to identify formats which presented difficulties in administration.

Long before June 1966, it became apparent that more time was needed for instrument development. From the very start, the emphasis of the

[3]Tyler, Ralph W. Introduction. In E. L. Norris (Ed.), *National assessment of educational progress: science objectives*. Ann Arbor, Michigan: Committee on Assessing the Progress of Education, 1969.

[4]The term "exercise" is used throughout to distinguish National Assessment materials from standardized test items.

FIGURE 1-1
Time Chart of Exercise Development


1966
1967
1968
1969
Jan Mar May July Sept Nov Jan Mar May July Sept Nov Jan Mar May July Sept Nov Jan Mar May July Sept Nov
Developmental Activity
1. Contractors Develop Exercises
2. Lay Review — Offensiveness
3. Mail Reviews
4. Feasibility Study In-School
5. Feasibility Study Out-of-School
6. Contractors Revise Exercises
7. 90% Study
8. First Subject Review Conference
9. Mathematics Study
10. Second Subject Review Conference
11. Choices Study
12. Tryouts
13. Lay Review — Meaningfulness
14. TAC Review
15. USOE Review
16. Final Selection

program had been on the development of assessment materials of highest quality. Figure 1-1 shows the steps related to exercise development which did occur and the approximate time requirements for each. The first six stages were those originally planned and at their conclusion it had been hoped the exercises would be ready for use in the actual assessment. However, with each of the early reviews of exercises, problems were pointed up which had not been obvious earlier and which led to more research and development until all of the stages on Figure 1-1 had been completed. (Each of the steps on the table will be explained in detail later in the monograph.)

All 10 subject areas went through at least the second subject review conference (step #10, Fig. 1-1). However, work from that point on did not progress at the same rate for all 10 areas. This was due to differing acceptance (or rejection) rates which occurred at the respective review conferences. It became evident that some areas needed a great deal more work than others. From among the five areas which could have been ready (Citizenship, Literature, Science, Social Studies and Writing), Citizenship, Science and Writing were chosen for the first assessment in March 1969.

The Organization of the Monograph

In the chapters which follow, the reader will be shown in detail the steps which were taken in developing the exercises for National Assessment, from the original criteria given to contractors through the various research studies, subject matter, lay and other reviews. The order of the monograph is chronological to show how the plan for exercise development grew into a complicated plan and how needs for further reviews and refinements became evident as the work progressed.

Throughout the monograph, examples of exercises and comments on them will be presented in an attempt to illustrate the types of changes which occurred at various steps along the way.

The reader should note that the examples used in this monograph are, for the most part, ones which have been dropped from the active National Assessment pools due to one or more major flaws. They are used here to demonstrate certain very specific characteristics of exercises and if evaluated as a whole may lead the reader to be justly critical. They *do not* reflect the quality of materials in active National Assessment pools. To use materials from active pools in a publication such as this would invalidate them for later use in the assessment.

Chapter II gives the rationale and criteria for the writing of the exercises. It explains that the first step in exercise development was the establishment of educational objectives for each of the subject areas chosen for assessment, and then specifies how the exercises were to be developed to sample these objectives. The same educational test development agencies which drew up the objectives were given contracts to write the exercises. Criteria relating to each of the following topics were established (to be explained in detail in Chapter II): content validity, clarity, variety of format, minimum of clustering, definite directionality, three difficulty levels, overlap between ages, special concern for out-of-school 17-year-olds and a need to sample content of all the objectives for a subject.

Special problems in relation to these criteria which arose during the development of exercises are also discussed. The whole area of attitude measurement and the difficulty of developing directional exercises which would measure progress over time was one of the big problems. Also, the need for developing scoring rationales and keys at the same time as the exercises are developed is currently stressed by National Assessment guidelines. Many of the ideas for imaginative exercises proved difficult to score objectively.

Chapter III takes up the early reviews of the exercises. One review given each exercise was to determine whether it would offend any large segment of the population. Lay people interested in education were given the responsibility of indicating which exercises were too offensive to be used. The other review was done by subject matter specialists by mail. These specialists reviewed the exercises to determine their content validity, whether or not there were obvious flaws in the exercise and whether the difficulty level given for the exercise seemed appropriate.

Chapter IV discusses the initial research studies that were done. One study was to determine whether or not the exercises written to be "easy" were, in fact, easy. The purpose of the other study was to look for problems in administering difficult or complex exercises to low ability students.

The reviews and studies discussed in Chapters III and IV are those originally proposed for the developmental work on exercises. However, it soon was apparent that more reviews and studies needed to be made. Chapter V considers the two sets of subject matter review conferences which were held and some of the special problems which arose in different subject areas. Chapter V also points out the need which arose for additional authors of exercises to fill in gaps in the coverage of objectives.

Three studies are discussed in Chapter VI. The Mathematics Study evaluated formal mathematical wording as compared with the use of less

precise language, and the value of using the "I don't know" choice on multiple-choice exercises. The Choices Study investigated the relationship between multiple-choice and open-end format. The final study reported was that of the actual exercise tryouts to determine whether any problems in administration still remained with the revised exercises.

Chapter VII describes the final reviews and the actual selection of the exercises to be used in the assessment. A conference of lay people was held to go over the exercises to determine that each exercise was asking for meaningful information. A final staff review of exercises was made by the Technical Advisory Committee (TAC)[5] for content validity, meaningfulness and clarity of wording. The exercises were then reviewed by the USOE to make a final judgment on whether any of the questions might be interpreted as an invasion of privacy. At the conclusion of these three final reviews, the exercises were deemed ready for selection for use in the assessment.

The final chapter tells what effect the first four years of research and development have had on National Assessment and indicates future directions in exercise development.

Example of Exercise

The following example of an exercise will illustrate the rather dramatic changes which sometimes occurred during the extensive refinement process. Not all exercises were affected by all of the reviews, but almost all

[5]The Technical Advisory Committee for National Assessment has been very active and influential in guiding the technical aspects of the project. Members of TAC include: Robert P. Abelson, Professor of Psychology, Yale University; Lee J. Cronbach, Vida Jacks Professor of Education, Stanford University; Lyle V. Jones, Director, The L. L. Thurstone Psychometric Laboratory, University of North Carolina; and John W. Tukey, Chairman, Department of Statistics, Princeton University. (Dr. Cronbach has not been an official member of the committee since January 1969, but has continued as a consultant.)

TAC was the primary source of guidance on all technical problems from the inception of the project until the operational phase began (1969), when the problems and needs for expert counsel increased so markedly that additional advisory groups were formed. TAC became the Analysis Advisory Committee (ANAC) and focused on problems related to the selection of exercises to be reported and the analysis of the data. In addition, the following advisory groups were formed:

1. Operations Advisory Committee (OPAC) to give advice on problems related to field operations,
2. Exercise Development Advisory Group (EDAG) to help guide further work in exercise development, and
3. Media Advisory Group (MAG) to give advice on matters related to reporting results.

of the exercises went through some editing as a result of one or another review.

The example is a citizenship exercise which was developed for ages 13, 17 and adult to sample the extent to which a person informs himself about the law, especially (for ages 13 and 17) in relation to his own status as a minor.

The original exercise read as follows:

Which of the following are allowed by law?

Allowed	Not Allowed	
_____	_____	Putting a person under 18 in jail.
_____	_____	Putting someone in jail without telling him why.
_____	_____	Putting someone who is in jail on a diet of bread and water.
_____	_____	Putting someone in jail if he hasn't enough money to pay a fine.

The lay panels which reviewed this exercise for offensiveness in September 1966, did not find it offensive but questioned the validity of the first answer. They recommended the use of the term "in custody" in place of "jail" as it would include juvenile court. In fact, they recommended that the exercise receive a thorough legal review.

A subject matter specialist who reviewed the exercise by mail reported that the exercise sampled the objective for which it was written, but he predicted it could be a very difficult exercise.

The exercise was returned to the contractor who revised it as follows:

Which of the following are allowed by law, and which are *not* allowed by law? (Mark X's to show which are allowed and which are not allowed.)

Allowed	Not Allowed	
_____	_____	Sentencing a person under 18 to state prison.
_____	_____	Putting someone in prison without telling him why.
_____	_____	Putting someone who is in prison on a diet of bread and water.

At the first subject review conference in the summer of 1967, the re-

viewers recommended that more choices be given, and they suggested the following be added:

_____	_____	Criticizing the government in a magazine.
_____	_____	Police entering your house without a search warrant.
_____	_____	Starting a new religion.

These three choices were added to the exercise before it was submitted to the second subject review conference in November 1967. The reviewers for age 17 liked the exercise but recommended the deletion of the first choice, "Sentencing a person under 18 to state prison." The panel for age 13 wanted extensive revision of the exercise and recommended a new exercise be written which would be more applicable to 13-year-olds. Their revised exercise was as follows:

Which of the following rights are protected by the U.S. Constitution? (Mark the boxes to show which are protected and which are not protected.)

Protected	Not Protected	
☐	☐	Carrying a concealed weapon
☐	☐	Causing a riot
☐	☐	Collecting evidence from a telephone conversation
☐	☐	Joining with others to promote a cause
☐	☐	Believing in God

The reviewers also wrote a new exercise they recommended for 13-year-olds:

Which of the following are allowed by law and which are not allowed by law? (Mark the boxes to show which are allowed and which are not allowed.)

Allowed	Not Allowed	
☐	☐	Throwing litter on the street
☐	☐	Driving as fast as you wish on the highway

☐	☐	Walking on private property
☐	☐	Entering a vacant house
☐	☐	Making people save for old age by taxing them

Both exercises were returned to the contractor for further work. The contractor dropped the old version in favor of the new version. The contractor had the new version reviewed by a group of students from low socio-educational backgrounds and asked them for their ideas. The students did not understand what was actually meant by the last choice, "Making people save for old age by taxing them." None of them connected this statement with Social Security.

The contractor resubmitted the exercise in the new form recommended by the subject review panel for 13-year-olds and kept in the last choice referring to old age. At this time an "I don't know" choice was added.

When the exercise in its final form was reviewed by the TAC, it was shelved[6] for the first round of the assessment as they thought it still needed more work.

[6]An exercise which is "shelved" is one which is withdrawn from the active pool and may receive reconsideration at a later date.

CHAPTER II

Rationale and Criteria for Writing Exercises For National Assessment

As a basis for common understanding of the purposes and developmental processes for the assessment instruments, a seminar lasting several days was held at the Center for Advanced Study in the Behavioral Sciences, Stanford, in the late winter of 1964-65. At that seminar representatives from interested contractors were present as well as National Assessment staff. The steps that were to be followed were explained and agreed to then.

The plan called for two major steps: (1) the development of objectives and (2) the development of exercises.

Objectives: Purpose and Process

Tyler has explained the purpose of objectives as follows:

> The major purpose for obtaining a list of educational objectives is to guide the development of the assessment instruments. The assessment should indicate the extent to which our people have attained important educational goals, i.e., have they learned what schools are expected to teach? Learning is a process of acquiring ways of thinking, feeling and acting, that is, acquiring patterns of behavior. A course outline commonly lists the content the course deals with but it does not indicate what the student is to learn to do with this content. Is he to memorize it, recognize the principles involved and seek to use the principles in solving problems, or to develop a skill in reading and analyzing the material or to develop a life-long interest in the subject, or some combination of these and other kinds of behaviors? We need to know what the student is expected to learn in order to assess his achievement. For this reason, educators are writing objectives in terms of the behavior they expect to help students learn. Objectives written in these terms are necessary to guide the production of assessment exercises.[7]

[7]Tyler, Ralph W. Personal letter, February 1, 1970.

In keeping with the goals of the project the objectives written for National Assessment have to be acceptable to three groups of people:

1. *Subject matter specialists.* Specialists in the subject area must consider the objectives authentic from the viewpoint of the discipline. Scientists must agree the science objectives are authentic; mathematicians must agree upon the authenticity of the mathematics objectives, etc.
2. *Educators.* School people must recognize them as desirable goals for education and ones which schools are actively striving to achieve.
3. *Citizens.* Parents and others interested in education must agree the objectives are important for youth and young adults to know, feel or understand.

These are very stringent criteria. It is not always possible to obtain complete agreement among these three groups or even within a group, but some consensus can be reached.

Within this framework the objectives were initially developed by different agencies under contract to National Assessment. Principals from all of the organizations thought to have capabilities of producing some of the needed instruments were contacted, and the assessment project was explained to them. Subsequently, bids were received and accepted from the American Institutes for Research (AIR), Educational Testing Service (ETS), Science Research Associates (SRA) and The Psychological Corporation (PC). The methods they used varied a bit, but generally followed the same procedures. The literature was surveyed to see what other groups had done, and then subject matter specialists and other qualified professionals were brought together to evaluate, to expand, to elaborate, to edit, to give direction to the contractor. From the literature, from their own resources and from consultants' suggestions the contractors produced the specific objectives of the Assessment for each of the 10 areas.[8]

When the contracting agencies delivered the objectives to National Assessment, an additional review was undertaken — a review by lay adults who were knowledgeable about education. Eleven panels of lay persons

[8]The objectives for five of the subject areas were developed by two different contractors. Having two contractors independently write objectives for the same subject area appeared to be a good way to find different approaches. Eventually a choice was made between the two sets of objectives for these subjects. The selected objectives then served as the basis for the production of exercises.

were organized representing four different parts of the country as well as cities, suburbs and rural areas. Participants were selected from nominations made by such organizations as the National Congress of Parents and Teachers, the National Association of State Boards of Education, the National School Boards Association and education committees of other organizations such as the National Association for the Advancement of Colored People and the National Association of Manufacturers. Each of the 11 panels reviewed the objectives in all 10 areas. All of the chairmen then met to consolidate the recommendations of the 11 panels. The result was that these lay panels accepted most of the objectives, but not all. They suggested editorial changes in a number of instances, and for one subject area they asked for a complete revision. The original objectives for the social studies area appeared to be lacking in clarity of direction for instrument development. The social studies objectives were reformulated in light of this suggestion.

The involvement of lay persons in the review of objectives is stressed here not because it was more important or more extensive than the reviews by subject matter specialists and other educators, but simply because it was a step that is not commonly undertaken by educators.

An additional point should be made about the nature of behavioral objectives as designed for National Assessment. The behavioral statements written for the general objectives are not exhaustive or completely definitive and therefore may not fit the definition of current proponents of behavioral objectives (e.g., Mager[9]). However, they do serve to describe the kinds of behaviors which might be observed relative to a more general objective. They are illustrative of desirable behaviors and provide the framework through which it is possible to determine what proportion of our population does exhibit these behaviors.[10]

For example, in Citizenship a general objective is to support rights and freedoms of all individuals. A subobjective under this objective is to recognize instances of the proper exercise or denial of constitutional rights and liberties, including due process of law. For age 17 this is illustrated by the following:

> Given a concrete example of any of the following, they should recognize it as a denial of the constitutional right:

[9]Mager, Robert F. *Preparing instructional objectives.* Palo Alto, California: Fearon Publishers, 1962.

[10]Copies of National Assessment objectives are available at $1.00 each in seven of the 10 areas — Citizenship, Literature, Mathematics, Music, Reading, Science, Writing. 201A Huron Towers, 2222 Fuller Road, Ann Arbor, Michigan 48105.

— denial of voting privilege by intimidation or unfair test.
— censorship of the press, mass media and public speech.
— police interference with assembly in a public place to peacefully protest an injustice.
— illegal search, arrest, or detainment.

In Science a major objective is that a scientifically literate individual will have knowledge of the fundamental aspects of science. One subobjective relates to inquiry skills necessary to solve problems in science, specifically the ability to recognize scientific hypotheses. Within this category fall such abilities as the ability to recognize data pertinent to a problem and the ability to recognize the possibility of testing a hypothesis. The 9-year-old should grasp the idea of cause and effect to be able to recognize possible explanations for the things he observes. The 13-year-old should be cognizant of the concept of cause and effect, know the nature of a scientific hypothesis and be able to recognize a simple explanatory hypothesis.

National Assessment does not assume that once its objectives have been developed they can be used indefinitely. From the beginning of the project it has been the intent to review the objectives of each subject matter area each time that area is assessed. Already the objectives for the three areas to be covered in the first year of the Assessment have been reviewed and revised, where necessary. National Assessment itself may acquire expanded or even different goals over time, as it responds to ever increasing demands for information pertinent to the evaluation of American education.

The Development of Exercises

After the objectives had been developed, the next step in the process was to develop a set of exercises. The term "exercise" is deliberately used to distinguish National Assessment materials from standardized tests. Most standardized achievement tests are normative and seek to evaluate the individual with respect to some group of people. The goal of National Assessment, however, is to measure the skills, knowledges and attitudes of large groups of people and determine their level of attainment.

With normative tests, items are selected that will spread individuals over a wide range of achievement so they may be placed in rank order. For the purposes of National Assessment, how widely persons differ in their educational performance is a fact to be noted, but it is not the basis

or goal of measurement. If a desired knowledge or skill has been acquired by everyone so that there is no variability in response on this exercise, it is still a significant fact for assessment. This fundamental difference meant that a new set of criteria had to be developed to serve as guidelines for the people who were to develop the materials for National Assessment.

The initial sets of exercises were developed by the contractors who developed the objectives for the 10 areas. The criteria initially established for the development of the exercises as well as those which evolved during the early developmental work fall under the following headings:

1. Content Validity
2. Clarity
3. Type of Exercises and Format
4. Clustering
5. Directionality
6. Difficulty Level
7. Content Sampling and Number of Exercises
8. Overlap Between Ages
9. Special Concerns for Age 17

Each of these is discussed more fully in the following pages.

1. *Content Validity*

The most important criterion which was established for exercise development was that every exercise must be a direct measure of some knowledge, skill or attitude which was stated in the objectives. That is, it must have content validity. An exercise has content validity if it is a direct measure of some important bit of knowledge, skill or attitude that reflects one or more objectives of a subject area. An exercise must be meaningful, make sense and be directly related to the objective. It must not be trivial, inconsequential or peripheral to the objective. In practice, then, an exercise has content validity if it makes sense to an informed reader who sees it together with an objective and says, "Yes, this a good measure of the knowledge or skill called for by this objective." In the review processes which were to follow, a reviewer read the exercise along with the objective to see whether the exercise was an appropriate measure for that objective.

For example, in Science the following exercise written to see if students are able "to check the logical consistency of hypotheses with relevant

laws, facts, observations or experiments," was judged appropriate by the reviewers:

In a particular meadow there are many rabbits that eat the grass. There are also many hawks that eat the rabbits. Last year a disease broke out among the rabbits and a great number of them died. Which of the following probably then occurred?

- ⬭ The grass died and the hawk population decreased.
- ⬭ The grass died and the hawk population increased.
- ⬭ The grass grew taller and the hawk population decreased.
- ⬭ The grass grew taller and the hawk population increased.
- ⬭ Neither the grass nor the hawks were affected by the death of the rabbits.
- ⬭ I don't know.

Relevance was also stressed. That is, the stimulus material should be meaningful to the person in terms of his day-to-day living whenever possible. For example, in Reading, the adult may be asked to read and interpret such things as directions for preparing income tax returns, telephone directories or parking tickets in preference to asking him to read and interpret a passage from a literary work he may not have seen since high school.

The emphasis on content validity and relevance points up one of the distinguishing features of National Assessment as compared with other "testing" programs. Current plans of National Assessment are to report the specific exercises which are asked, together with details regarding right and wrong responses. For multiple-choice type exercises, the number and percent of responses made to each choice will be reported. For open-end exercises, actual samples of both correct and incorrect responses will be reported. For essay-type exercises, actual sample responses will be reported. This procedure contrasts sharply with the usual practice in reporting results of standardized tests. In the usual standardized test, individual items are not generally reported or revealed — either to the individual taking the test, or to school boards or other groups interested in evaluating educational attainments. Revealing the actual items destroys the usefulness of a test unless the revealed items can be replaced immediately by new ones. However, National Assessment's reporting plans do call for reporting individual exercises and replacement with new ones. Since a major purpose of the program is to report to the lay public as well as educators, it became of major concern that the relevance of the exercises to the objective be clearly apparent and the content be meaningful.

Extensive review processes by both subject matter and lay groups became the chief vehicle for determining the content validity of the exercises. Groups of scholars, educators and lay persons representing widely diverse geographical areas and points of view have reviewed each exercise and made recommendations to accept, reject or change the materials.

2. *Clarity*

From the beginning, emphasis was placed on making each exercise understandable. The student must first know what he is being asked to do. Too often the vocabulary level of the traditional test item and the reading disability of the student prevent the student from demonstrating the capabilities he possesses.

Writers were urged to state all exercises as simply and directly as possible so as to maximize understanding and minimize reading disability.

As work progressed, other major decisions were made to alleviate the heavy demands usually made on reading and writing skills when assessing any area:

a. Group administered exercises would be tape administered and each question would be read to the student (except in the assessment of Reading).
b. Exercises which require lengthly written responses (except in the assessment of Writing) would be individually administered and the administrator would either tape the response or write it down for the respondent. (Subsequently many of the 9-year-old exercises requiring even a short response were designated for individual administration.)

3. *Type of Exercises and Format*

Writers were encouraged to use that format or mode which could provide the best and most direct measure of the objective being assessed. For example, to assess a student's ability to handle scientific apparatus he should be given the apparatus to handle, not a picture of it. Wherever possible, writers were encouraged to develop exercises which would obtain actual samples of the student's skill in an area. Some areas such as Writing, Art and Music lend themselves readily to the development of such performance measures.

In assessing knowledges and attitudes, writers were encouraged to break away from the traditional testing approach which relies heavily upon group paper and pencil tests and easily scored multiple-choice type exercises. If a better measure could be obtained by using an individual

interview technique, it should be preferred to the more widely used group testing situation. Open-end exercises which required the student to recall and apply information were to be preferred over multiple-choice forms which only require the student to recognize and select the correct answer from the set of choices. The use of new and unique formats was encouraged, provided they could be readily understood by the student.

As a result of this encouragement, the contractor for Social Studies proposed using film strips and asking questions about the strips afterwards. One exercise of this type was designed to assess a person's belief in the freedom of the first amendment. (This exercise was temporarily shelved because it needed further work.) The exercise is as follows:

> You are now going to see the film *Censorship: A Question of Judgment*. Watch the film and then answer the following questions about the film.
>
> (a) Explain briefly what Nancy and Mr. Bishop are disagreeing about.
>
> _____
>
> (b) Do you think the picture of the fight should be printed in the school newspaper?
> ☐ Yes
> ☐ No
> ☐ Undecided
> Explain the reason for your answer.
>
> _____
>
> (c) Who should decide what is printed in a school newspaper?
> ☐ Student reporters and editors
> ☐ The teacher who acts as sponsor
> Other (Please state)
>
> _____
>
> (d) Do you believe there are certain areas (for example, movies, books, television) where there should be some form of censorship?
> ☐ Yes
> ☐ No
> ☐ I don't know.
> Briefly explain the reasons for the answer you marked above.
>
> _____

In the area of Science, some exercises involve the use of actual laboratory apparatus. In Art, many pictures are being used. In Music, tapes of music are being played for some exercises; in others the student is recorded on a tape while he sings or plays a musical instrument. In Citizen-

ship, groups of young people are brought together to discuss vital issues, and they are rated on their behavior in a group situation. The liberal use of apparatus, diagrams, pictures, filmstrips and other media was also encouraged.

This "wide-open" approach was only partially successful. Writers differ in their talents and degree of flexibility. Some were able to break away from the traditional stereotyped way of writing test items, others were not. When materials did not reflect the qualities which National Assessment sought, contracts were either renegotiated or independent writers were commissioned to improve the variety and quality of exercises. In addition, special problems relating to scoring and formats were encountered.

Scoring Problems. Problems were created in this "wide-open" approach, not the least of which was the development of adequate scoring rationales and keys for non-multiple-choice type exercises. When exercises were first written, only very broad, general scoring schemes were presented. Many of these did not adequately communicate to staff, reviewers and other contractors exactly how an exercise was to be scored.

Some writers felt that their exercises could only be scored after they were administered, and then the writer could categorize and define correct and incorrect responses or rank them on a scale. This approach did not prove to be feasible. It created problems at subject review conferences since a vital part of evaluating an exercise is related to knowing how it will be scored. It created similar problems at the time exercises were selected to be used in the first assessment when some materials were rejected which might have been used had the scoring been better defined.

As a result of these experiences, the early developmental process now stresses the detailed definition of how each exercise is to be scored. A technique similar to that used by individually administered tests such as the Stanford-Binet and the Wechsler scales is being developed. That is, specific student responses are obtained through preliminary tryouts and categorized as acceptable, unacceptable or questionable for each scoring category for every open-end exercise. In the case of exercises which need to be ranked, such as Writing, actual sample responses are obtained in the preliminary developmental stages so that scoring rationales give illustrations of the quality of response typical at different points on the scale. This is a time consuming and costly process, but early and close monitoring of exercises with emphasis on scoring improves quality significantly.

Format Problems. Another problem created by the "wide-open" ap-

proach was related to format of the paper-and-pencil exercises. Many new and interesting formats did evolve among the paper-and-pencil exercises. However, it was found through special studies that some formats were too complicated and tended to confuse the task. For example:

Directions: Choose the best answer.
Which of the following do all cultures have?
I. Some form of education
II. Incest taboos
III. Rule by kings
IV. A belief in one God

(A) IV only
(B) I and II only
(C) III and IV only
(D) I, III, and IV only

Then, too, each new format had to be accompanied by detailed instructions on how to mark the exercises. Much of the time allotted for administration was spent in trying to communicate these instructions.

As concern over formats developed, it became necessary to catalog the various types of formats and to introduce some uniformity into the way each type was to be presented. In the summer of 1967, AIR was commissioned to examine the exercise formats in the current pools and to prepare a document describing the basic kinds of formats then in use, together with examples to serve as guides for the future.

AIR's analysis described nine basic formats then in the various subject pools:

1. Completing —
This type requires the student to fill in blank(s), or to provide a short answer following some stimulus or to arrange alternatives in correct locations. For example:

What are the capital cities of the following states?
New Hampshire _____
Oregon _____
Texas _____
Georgia _____

2. Writing —
The response to this type requires the student to write a sentence(s), or paragraph(s) or list responses. For example:

Write a set of directions explaining how to get, by car or on foot, to some famous local landmark from the airport, railroad station, bus station, or turnpike (freeway) exit closest to where you live. For instance, someone living in Trenton, New Jersey, might explain how to reach the place where George Washington crossed the Delaware River. Someone living in San Antonio, Texas, might explain how to get to the Alamo. Someone from the northern part of New York State might tell how to get to Niagara Falls or the Baseball Hall of Fame at Cooperstown.
Write your directions carefully and clearly, as if you were going to give them to a friend who is not very familiar with your area.

_____

_____

_____

3. Checking —
Two or more response columns must be checked against a stimulus list. For example:

Indicate whether each statement below is a fact *or* an opinion by marking in the appropriate box to the left of the statement.

Fact	Opinion	
☐	☐	John F. Kennedy was elected President of the United States in 1960.
☐	☐	I think it snowed on election day in 1960.
☐	☐	Kennedy was opposed by the Vice-President, Richard Nixon.
☐	☐	Lyndon B. Johnson was Kennedy's running mate.
☐	☐	If it had not been for the television debates, Nixon would undoubtedly have defeated Kennedy.
☐	☐	John F. Kennedy was the best President we ever had.

4. Checking a list —
This format requires the student to check or mark one or more responses from a single column list. For example:

If a person likes work of a mechanical nature, which of the following jobs could *best* satisfy his interest? (Check as many as apply.)

☐ Layout man
☐ Actuary
☐ Chef

☐ Shipping clerk
☐ Surveyor
☐ Millwright
☐ Economist
☐ Barber
☐ Printer
☐ Farmer

5. Ordering —

This format requires the student to number some or all alternatives in sequence, or to arrange alternatives in correct sequence. For example:

You are to choose the five types of reading matter you like best from the list below. Decide which kind of reading you like best and put a number 1 in the space in front of that type. Then decide which kind of reading you like second best and write a number 2 in the space in front of it. Show which kinds of reading you like third, fourth, and fifth best by putting a 3 in front of your third choice, a 4 in front of your fourth choice, and a 5 in front of your fifth choice.

____ Comic books
____ Poetry
____ Adventure stories
____ Mystery and detective stories
____ Fiction
____ Science fiction
____ Sports stories
____ Plays
____ Biography and autobiography
____ Mathematics
____ Art
____ History

This particular exercise lacks directionality (see Section 5) but is used here to illustrate the ordering format.

6. Matching —

The matching format requires the student to pair stimulus and response alternatives. For example:

Here are two lists. One list is of music. The other list is of composers who wrote the music. Write the letter of the composer in the blank next to the music he wrote. The letter of each composer may be used once, more than once, or not at all.

Music	*Composer*
____ "Peter and the Wolf"	A. Ludwig van Beethoven
____ "Carnival of the Animals"	B. Leonard Bernstein
____ "The Nutcracker Suite"	C. Aaron Copland

_____ "The 1812 Overture"	D. Edvard Grieg
_____ "Stars and Stripes Forever"	E. Wolfgang Amadeous Mozart
_____ "William Tell Overture"	F. Sergi Prokofiev
_____ "Blue Danube Waltz"	G. Giacchino Rossini
_____ "The Firebird"	H. Camille Saint-Saëns
	I. John Philip Sousa
	J. Johann Strauss
	K. Igor Stravinsky
	L. Peter Ilich Tchaikovsky

7. Marking —
A response is made on a diagram or picture, word or sentence. For example:

In the figure shown below, every row and every column will add to the same total if the number in *one* of the nine squares is changed. Find this number and put a circle around it.

12	23	15
17	14	19
21	18	16

8. Multiple-choice —
The correct response is selected from two or more alternatives. For example:

Directions: Mark only *one* answer.
John counted his breathing rate several times during two days. His record is shown below.

BREATHING RATES

	Time of Day		
Day	*Morning*	*Noon*	*Night*
Tuesday	19	16	14
Wednesday	20	16	13

When was John's breathing rate highest?

- ☐ Tuesday in the morning
- ☐ Tuesday at noon
- ☐ Tuesday at night
- ☐ Wednesday in the morning
- ☐ Wednesday at night

9. Combination —
The combination form is a multiple-choice exercise followed by either completion or writing. For example:

Directions: Mark your answer to the first question. If you marked "yes", answer the second question briefly.

Have you ever been in the city where your state capital is located?

☐ yes

☐ no

(If yes) What one thing that you saw in that city do you remember most?

_____

_____

_____

Upon later recommendation of the Technical Advisory Committee (TAC), the existing exercises using the checking format were changed to multiple-choice types. The multiple-choice format seemed to be more easily understandable by low achieving students.

The checklist format also proved to be undesirable because of problems created in trying to score those exercises in which there was more than one right answer. For example, would three correct answers plus three incorrect answers be judged equivalent to three correct answers and no incorrect answers?

In the first assessment only four of the remaining formats were actually used: (1) completion, (2) writing, (3) multiple-choice and (4) combination. Because of the previously mentioned problem of communicating detailed instructions to the respondent, future assessments will probably be somewhat limited in the number of different formats which can be used, though not necessarily limited to these four types.

Another kind of format consideration arose specific to multiple-choice type exercises. Such questions as these were considered:

1. Should the alternatives be lettered or numbered?
2. Should the student be required to check (√), mark an X, or fill in a box (☐) or use some other method to indicate his answer?

Both questions are related to how the exercises are to be scored. If exercises are to be "machine read," it is only necessary to have a mark in a specifically designated place on the paper. Numbering or lettering the alternatives is not necessary. However, if the scoring process involves

key punching, a number or letter is required to insure accuracy by the key punch operator. How the student marks (with a check, X or other method) does not matter.

This particular issue was not settled until the scoring contract was awarded for the first assessment. The Measurement Research Center (MRC) in Iowa City was awarded the contract for scoring, and the multiple-choice exercises were designed to fit their specifications for machine scoring using optical scanning equipment and the alternatives were not lettered or numbered.

4. *Clustering*

Original criteria did not specify rules about clustering. A cluster consists of a series of questions based on one set of stimulus material. Some writers argue that it is inefficient to ask only one question when the length of the stimulus requires the student to spend 10 or 15 minutes just to understand it. Another argument for clustering is that some exercises have little or no significance when asked by themselves, but when grouped with other exercises take on significance. One such example of the latter from the field of Literature is an exercise in which students are asked to identify illustrations from specific literary works. For a student to be able to identify "Snow White and the Seven Dwarfs" may not in itself be a thing of great importance, but the student's ability to identify at least three out of 10 literary works from their illustrations does take on significance.

Arguments against clustering relate to the relative weight which is to be given to a specific objective. There will always be some limitations on the total number of exercises which can be used in an assessment. In the first assessment each subject area was allotted an average of 160 minutes per age level, due primarily to financial considerations. While this particular time limit may be changed in future years, it is reasonable to expect there will continue to be a defined amount of time which can be spent in assessing a given subject area. The question becomes a relative one. If a 30-minute clustered exercise is used, it necessarily takes time from the assessment of other objectives for that subject. More than one long clustered exercise in an area rapidly uses up the allotted time. Maintaining adequate coverage of all objectives is essential to the assessment.

A compromise has been reached in that the total amount of time required to administer the exercise is now the governing factor. It makes little sense to develop many long clustered exercises because this limits the other materials which can be used. On the other hand, it does seem

inefficient to use only one question if the stimulus is unusually long. Clusters of usually not more than three or four parts are allowed, provided the time is not excessive and there is a good reason for grouping.

5. *Directionality*

The criterion of directionality is related to the measurement of progress over time. In order to measure progress, it is necessary to take measurements at two different times and to accept change in a specific direction as progress.

Each exercise developed for National Assessment must have a right or wrong answer or be assigned some value on a scale. Since the objectives represent desirable achievements, the exercises are designed to assess the proportion of our population of a given age exhibiting that achievement. Thus, all exercises must be scorable in a desired direction. That is, an exercise must have directionality. There must be agreement upon the desired direction of response.

In the cognitive areas this requirement was fairly easily met. Knowledges and skills are generally definable in terms of correctness. However, this was not so in the area of attitudinal assessment. Many attitudinal exercises simply do not have a correct answer or a desired direction.

An example of an exercise which lacks directionality is the following one taken from Literature. The exercise was developed to assess: (1) recognition of the importance of literature to the individual and society, (2) recognition that literary expression requires a number of forms in order to enable it to become art and (3) recognition of the necessity of a free literature in a free society. The exercise is as follows:

> People have often argued that certain types of books and movies should be kept from being sold to the general public. What is your opinion with respect to this matter? Please fill in the box next to the letter of the opinion that is closest to yours.
>
> Do you think that a book that is unpatriotic
>
> ☐ (A) should be reviewed by a board of review and banned for everyone if it seems offensive?
>
> ☐ (B) should be reviewed by a board of review and banned for minors but not adults if it seems offensive?
>
> ☐ (C) should be reviewed by the head of a family and banned for minors of that household if it seems offensive?
>
> ☐ (D) should not be restricted in any way?
>
> ☐ (E) no opinion

This exercise was judged to lack directionality because it was ambiguous. There was no agreement on what the "correct" answer was. Widely differing views are held on censorship and no one was willing to say that there was one acceptable correct response to this exercise.

Another literature exercise which lacks direction is the following one which was developed to assess whether an individual has developed a continuing interest and participation in literature and the literary experience:

> This is NOT a test. We would simply like to find out what sorts of books you, and other people like you, enjoy reading. Because we cannot show you all the different kinds of books, we have to use a list. This list gives a brief description of what each kind of book is about. In front of each of the descriptions are two words, *YES* and *NO*, and a question mark *(?)*. If you would like to read the book, circle the word *YES* in front of the description. If you would not like to read the book, circle the word *NO*. If you cannot make up your mind, circle the question mark. After you have gone through the list, please answer questions 40 and 41. (Only a few of the 39 choices are listed below.)
>
> 18. YES NO ? A story about a career woman in New York and her romantic adventures.
>
> 25. YES NO ? A story in which a middle-aged man discovers that he failed in his career and his marriage.
>
> 27. YES NO ? A play in poetry about how a young man tricks a whole town and marries the mayor's daughter.
>
> 38. YES NO ? A guide to manners, good grooming, and how to be more attractive to men.
>
> 40. Please write the title of your favorite book on the line below.
>
> _____
>
> 41. If you could tell a writer to write a book just for you, what sort of a book would it be? What would it be about?
>
> _____

A questionnaire exercise like this one does not assess progress over time in the participation of individuals in the literary experience. How could the 39 different types of books be compared? Are some choices preferable to others? Could it really be shown that over a period of time one group of students in answering this exercise showed more interest in

literature than another group had shown? This particular exercise is also a good example of "clustering" which was mentioned in an earlier section of the chapter. Thirty-nine of its 41 items are dependent upon the same set of directions. However, it would be difficult to de-cluster it, as choosing among only four or five sample book plots would indicate very little about a person's "participation in the literary experience."

The measurement of attitudes often falls more easily into the realm of surveys, where the purpose is to obtain information and observe change without any attempt to place a value judgment on the desirability or desirable direction of change. This does not mean that exercises of an attitudinal nature cannot have directionality. It has simply proved more difficult to define and reach agreement on what the desired direction of change is. The measurement of attitudes has remained one of the most difficult, and as yet unresolved, problems in National Assessment.

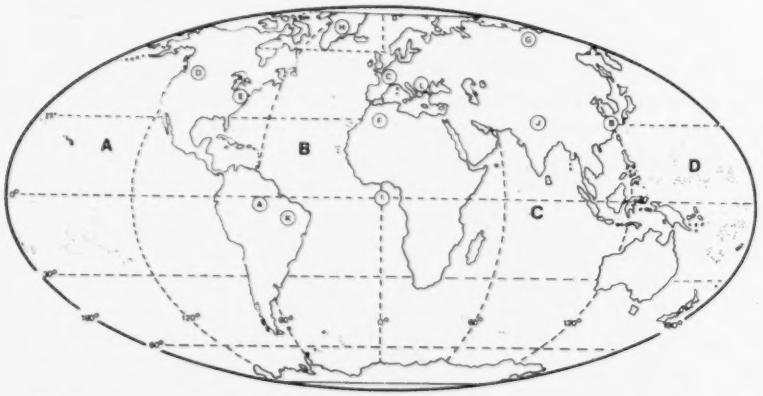
6. *Difficulty Level*

How easy or difficult an exercise is is judged by the percentage of people who can successfully answer it. An exercise which can be passed by 80 percent or 90 percent of the people taking it would be an "easy" exercise, while one which could be passed by only 10 percent or 20 percent of the people would be a "difficult" one. Most standardized tests are constructed in such a way as to facilitate the ranking of students from highest achieving to lowest achieving. This is done by designing most exercises so that they can be answered correctly by about one-half of those tested (about 40 percent to 60 percent) and includes just a few very easy or very difficult exercises.

Since National Assessment does not intend to place people in rank order, making most exercises of medium difficulty level is not appropriate. The intent of National Assessment is, instead, to be able to describe the knowledges, skills and attitudes of the most able and the least able students at a given age in addition to those of the average or typical students. Thus equal numbers of exercises had to be aimed at each of these three groups. For convenience these exercises have become referred to as 90 percent, 50 percent or 10 percent exercises, although it is understood that they actually represent exercises in that general range. A few examples will illustrate the point.

An example of an easy or 90 percent exercise is the following one written for 17-year-olds to assess the knowledge of spatial distribution of significant elements of the physical environment, for example: water.

Directions: Mark the correct answer.



On the map above, which letter represents the Atlantic Ocean?

- ☐ A
- ☐ B
- ☐ C
- ☐ D
- ☐ I don't know.

An example of an exercise of average (50 percent) difficulty is the following one written for adults to assess their knowledge of the diversity of living things:

Directions: Choose the best answer.
Which of the following is an animal?

- ☐ (A) Bacterium
- ☐ (B) Sponge
- ☐ (C) Toadstool
- ☐ (D) Snapdragon
- ☐ (E) Moss

A difficult, or 10 percent exercise, is exemplified by the following one written for 13-year-olds to assess their understanding of some of the characteristics of American government:

Who is the commander-in-chief of the United States armed forces?

- ☐ (A) The Commandant of the Marine Corps
- ☐ (B) The Chief of Naval Operations
- ☐ (C) The Secretary of Defense
- ☐ (D) The President
- ☐ (X) I don't know.

It has been argued that if the materials are able to be developed in this way (equal numbers of 90 percent, 50 percent and 10 percent exercises), an assessment does not need to be made because we would know the results in advance. However, the three designations are only "target areas." If the low achieving population is to be assessed, there must be *some* exercises on which they *can* succeed. When an exercise writer creates an exercise he "estimates" what percentage of an age group can answer it correctly. Field trials will help improve this estimate, but field trials rarely represent the conditions of actual assessment. Hence materials aimed at 90 percent, 50 percent and 10 percent levels of difficulty are only approximations at the time they are used in the assessment.

As it turned out, the most difficult part of these criteria proved to be the production of adequate numbers of "easy" exercises. Writers, when "arm-chairing" the difficulty level of exercises, tend to think an exercise is easier than it actually is. Since writers are not normally called upon to write large numbers of easy exercises, this finding was not too surprising. The details of this particular problem (the 90 Percent Study) are presented in Chapter IV.

7. *Content Sampling and Number of Exercises*

The objectives had been developed with the idea that if a given behavior was a desired goal of education, it should be included. It was fully recognized that some desired outcomes of education would be difficult, or even impossible, to measure. However, the objectives were to serve as the guides to the production of exercises and the coverage was to be as complete as possible. When this decision was made, it was not known how many exercises or how much time would be allowed for each subject area in the assessment. It was generally recognized, however, that a pool should be large enough to allow for a good selection of exercises.

If materials were to be acceptable to subject matter specialists, educators and laymen, a substantial number of exercises could be expected to be lost through the necessary review processes. In anticipation of this potential problem, contractors were directed to prepare an overage of exercises. The loss which resulted from reviews varied widely from one subject area to another. In Art, for example, it was necessary to shelve much of the material because subject matter reviewers would not accept the exercises as valid measures of the objectives. Art provided one of the more extreme cases. Whenever objectives suffered loss of coverage due to the review process, additional exercises had to be written.

Another reason to produce more exercises than were needed for the assessment was so that some could be held for future years. Reporting plans call for revealing only part of the actual exercises each time a subject area is assessed; hence there is a need to replace those exercises.

The question of how many exercises to prepare was considered in the initial planning stages but was never actually pre-determined. There appeared to be too much variation from one subject area to another to attempt to standardize the number of exercises across all 10 areas. Some areas such as Writing could be expected to have fewer exercises, but ones which required more time to administer. Other areas such as Science or Mathematics could reasonably be expected to have a greater number of exercises, but shorter in length. Allowing each contractor to specify his plans for content sampling and number of exercises resulted in wide variation in amount of materials produced and indicated that some guidelines would be desirable.

There is general agreement that the objectives should be covered insofar as possible. However, just what constitutes adequate sampling is an issue which has not really been faced and resolved. Perhaps the objectives themselves must ultimately determine what adequate coverage is. In a field such as Science and Mathematics there is a great deal of content, and it is not realistic to think that all can be sampled in any given assessment. However, in the area of Writing, a product (the ability to write) is being evaluated. The number of contexts in which it is possible to evaluate Writing is more limited than the knowledges and understandings required in other subjects. Hence Writing, as it relates to the major contexts (social, business, scholastic), can more easily be covered in any given assessment than a subject with a huge content area.

Experience has shown that it is necessary to specify a maximum volume of exercises to be produced. There was wide variation in the sheer volume of materials originally produced by contractors (see Table 2-1). In the area of Mathematics the final pool before assessment was approximately 2,400 exercises for the four age levels, or an average of 600 per age level.[11] At an average of one minute per exercise the pool contained 10 hours worth of material per age level! While an overage was considered desirable, and while the total number of minutes which could be assessed was unknown in the early developmental stages, 10 hours clearly represented far more material than could be used.

[11]Additional exercises were written to give better coverage of the objectives after receiving an initial 1,911 from the contractor.

As it happened, the amount of time which could be devoted to assessing an area was governed in large part by financial considerations. In the first assessment this averaged 160 minutes per age for each of three areas. The allotment was specified in minutes rather than in number of exercises due to differences in objectives and the wide variation in types of exercise among the areas being assessed. For example, many of the writing exercises took 20 to 30 minutes to administer, while the majority of science exercises took only one minute to administer. It took fewer (in number), but longer (in time) exercises to cover the objectives in Writing, whereas the science objectives needed more (in number) but shorter (in time) exercises for adequate coverage. The 160 minutes is considered minimal by most subject matter people involved, and steps are being taken to relax this current restriction.

It is obvious that some limits need to be set on the production of exercises, and contractors are now asked to prepare not more than 500 minutes worth of exercises per age level, a number which still allows adequate overage provided the quality of the exercises is good.

TABLE 2-1

Approximate Number of Exercises By Age and Subject[a]
April 1967

Subject	9	13	17	A	Total
Art	484	533	726	725	2468
Career and Occupational Development[b]	111	569	1510	1469	3659
Citizenship	92	172	159	177	600
Literature	361	464	638	632	2095
Mathematics	585	403	587	336	1911
Music	321	394	634	600	1949
Reading	1350	1493	1291	1291	5425
Science	180	239	228	196	843
Social Studies	189	387	749	451	1776
Writing	72	70	81	60	283
Totals	3745	4724	6603	5937	21,009

[a]Tabulations include figures for all exercises including those overlapping exercises which are used at more than one age level. Counts changed sometimes radically as work progressed on the materials.

[b]Career and Occupational Development was originally called Vocational Education.

8. *Overlap Between Ages*

In National Assessment an exercise which "overlaps" is one that is appropriate for more than one age group. An exercise generally overlaps between two adjacent age groups, but may overlap three or even all four age groups, or two non-adjacent age groups (e.g., 13 and adult).

While not in the original specifications to contractors, the design of overlapping exercises was encouraged fairly early in the project's history. It was anticipated that some interesting comparisons could be made by giving the same exercise to two or more age groups. Normally, one would expect an increase in the percentage of students who could answer an exercise correctly with increasing age. However, it is possible for this to be reversed for certain age groups where specific knowledge is involved. It is possible, for example, for the adult group to perform less well than 17-year-olds due either to "forgetting" or to rapid changes in the curriculum.

9. *Special Concerns for Age 17*

While it was expected that most 17-year-olds would still be in school, it was recognized that there is also a population of out-of-school 17-year-olds. This out-of-school group may be composed of high school drop-outs or of bright advanced students who have gone on to college. Because the out-of-school 17-year-olds might be difficult to find in large enough numbers to assess in groups, it was readily seen that exercises for 17-year-olds had to be developed which were amenable to either group administration or to individual interview. For assistance in this matter, survey specialists were consulted. Work with out-of-school groups both in try-outs and research studies showed that a few minor changes were all that were necessary to adapt most exercises to either mode of administration.

Summary

Since the usual rules for test construction did not appear to be useful in developing exercises for National Assessment, a new rationale and new criteria had to be developed.

In developing objectives:

1. Subject matter specialists had to agree that the objectives were authentic to their respective disciplines.
2. Educators had to accept the objectives as goals which were important outcomes of education and ones toward which they were teaching.

3. Informed lay persons had to agree that the objectives were important goals for the youth of the country to be striving toward.

Major consideration in developing materials was given to:

1. Content validity — developing exercises which are appropriate measures of the objectives.
2. Clarity — developing exercises which are stated as simply and directly as possible.
3. Type of exercises and format — developing exercises using whatever mode provides the best and most direct measure of the objectives being assessed.
4. Clustering — grouping exercises only for very specific purposes and limiting the amount of time necessary to administer such exercises.
5. Directionality — developing only exercises which have a right or wrong answer or can be scored on a scale.
6. Difficulty level — developing roughly equal numbers of "easy," "average" and "difficult" exercises.
7. Content sampling and number of exercises — developing exercises which adequately cover the objectives and developing enough exercises so that some could be lost through review processes with enough remaining for the assessment.
8. Overlap between ages — developing exercises which are appropriate for use at more than one age level.
9. Special concerns for age 17 — developing exercises for 17-year-olds which can be either group administered in-school or individually administered out-of-school.

CHAPTER III

Initial Reviews

The original plan for the development of exercises called for a review by lay people and one by subject matter specialists, followed by a tryout of exercises to determine feasibility of administration. The purpose of the lay review was to identify exercises which might be considered offensive and should be eliminated for this reason. The review by subject matter specialists was to assure that each exercise was a valid one which did assess the educational objective for which it was written, was an exercise which represented the discipline accurately and was one which had no identifiable flaws. Hopefully these two reviews would be sufficient to identify all of the questionable exercises and the tryouts would successfully identify those exercises which presented difficulties in administration. Subsequent events showed that this was an overly-optimistic plan.

Review of Exercises for Offensiveness[12]

Concern about possible offensiveness of exercises leads to a unique approach in exercise development. Although publishers of standardized tests have occasionally found themselves accused by the public of asking offensive questions, little systematic effort has been made in the past to determine just what is offensive to the general public. In most cases, it has been the judgment of one person or at most a few people as to how the public would react to specific questions.

Concern about the importance of not including any excessively offensive exercises and taking positive steps to eliminate such materials proved its value as the assessment later got underway. Relatively few questions were objected to by participating schools and a high proportion of schools approached agreed to cooperate.

As stated earlier, one of the major responsibilities of National Assessment was to confer with teachers, administrators, school board members

[12]An article by Frances S. Berdie based on this material appears in the June 1970 issue of *American School Board Journal*.

and others concerned with education to obtain advice and recommendations on the ways in which a national assessment of educational progress could be designed to be constructive and useful. In carrying out part of this responsibility, the staff called upon lay people to review the 10 original subject area objectives. Lay persons actively interested in education were identified at that time by asking for nominations from the officers of various national and state organizations interested in education. As actual exercises were developed, these lay people were again consulted.

Purpose of Review

As exercises used in the assessment should not be offensive to the general public, the lay people were asked to review the exercises from the viewpoint of the questions: "Would you object to having the exercise used with your child?" and "In your opinion would any important group in your community object to the use of the exercise?" An exercise was to be considered offensive if an appreciable number of people would consider the question as an invasion of privacy or as offensive in some other manner. The judgments they were asked to make were more difficult than it would appear as there are considerable differences in what is considered offensive in different parts of the country and in different social strata. One of the problems the lay panels faced was to decide whether or not an exercise was sufficiently important to risk seriously offending certain segments of the population.

Method

Obviously the lay people could not be asked to review every one of the over 21,000 exercises which had been developed. As some of the exercises, such as ones in Mathematics, were not at all offensive, the staff screened the exercises to identify those which might be considered offensive. These exercises were then assembled and printed into five booklets and discussed at five lay conferences held during 1966 and 1967. People invited to attend each conference were chosen to give representation from each area of the country (Northeast, Central, Southeast and West), from different-sized communities (large city, suburban and rural),[13] from different races, religions, sexes and economic levels of community and from different organizations interested in education. The representation by region of the country is shown on Table 3-1. Forty-

[13]These community categories were later regrouped into four classifications: large city, urban fringe, middle-sized city and rural small town.

TABLE 3-1
Lay Participants on Exercise Review Panels
May 1966 - March 1967
Distribution by Region and State

Northeast		*Southeast*		*Central*		*West*	
No. People	State	No. People	State	No. People	State	No. People	State
1	Connecticut	4	Alabama	2	Kansas	1	Alaska
1	Delaware	3	Arkansas	7	Illinois	3	Arizona
2	Dist. of Col.	2	Florida	1	Indiana	9	California
1	Maine	3	Georgia	1	Michigan	2	Colorado
4	Massachusetts	1	Kentucky	6	Minnesota	1	Hawaii
3	New Jersey	1	Louisiana	6	Missouri	2	Idaho
10	New York	1	Mississippi	4	Nebraska	1	Montana
3	Pennsylvania	2	S. Carolina	1	N. Dakota	1	Nevada
2	Vermont	2	Tennessee	2	Ohio	1	New Mexico
		5	Virginia	3	Oklahoma	2	Oregon
				1	S. Dakota	3	Texas
				1	W. Virginia	4	Utah
				3	Wisconsin	5	Washington

	Total Number of	
	People	*States*
Northeast	27	9[a]
Southeast	24	10
Central	38	13
West	35	13
Grand Total	124	45

[a]Counting the District of Columbia

five of the 50 states were represented by a total of 124 people. Twenty-seven came from the Northeast, 24 from the Southeast, 38 from the Central and 35 from the West. The number of panelists from large city and rural areas was about equal, but slightly fewer people came from suburban school systems. The panelists included 77 men and 47 women. The organizations represented are listed in Table 3-2.

TABLE 3-2
Lay Participation on Exercise Review Panels
May 1966 - March 1967
Distribution by Organizations Interested in Education

Organization	
American Association of University Women	5
AFL-CIO	2
College Education Student	1
National Association for the Advancement of Colored People	2
National Citizens Committee for Support of Public Schools	1
National Conference of Christians and Jews	1
Parent Teacher Association	
National Congress of Parents and Teachers	6
State Parent and Teacher Organizations	11
School Boards	
Catholic Education Organizations	4
County Boards of Education	5
Local Boards of Education	16
National Association of State Boards of Education	2
National School Boards Association	11
State Boards of Education	6
State School Board Associations	23
U.S. Chamber of Commerce, Education Committee	4
Miscellaneous	24
Total	124

The number of people at the conferences varied from 17 to 46, and the participants were divided into smaller work panels averaging 10 people each. A participant chaired each panel and a staff member was present to serve as a resource person. At least two separate panels of reviewers were formed at each conference, and at one conference there

were four panels. Each conference discussed different exercises, but everyone at any one conference discussed the same exercises. Hence, every exercise received review by from two to four panels of reviewers. When so many different people and points of view are represented, conflicting recommendations are inevitable. Toward the close of each conference, the participants met together to talk over their reactions to the exercises. These discussions were frequently lively, as what was offensive to some people others thought very vital and important to include. Attempts were made at this time to iron out the differences, but if this was not possible, the reviewers were told that National Assessment would make the final decisions on what was to be included, excluded or changed. After each lay conference on exercises was concluded, the staff members who had acted as resource people in each panel pooled the ideas which had been expressed in their respective panels. If a majority of the panels felt that an exercise was potentially offensive, the staff usually decided to omit it. However, if it could be rewritten to alleviate the offensive quality and if it assessed material which was deemed important, it was returned to the contractor for revision. If only one panel had strong reactions to an exercise and the other panels found it unobjectionable, it was usually retained. This was not always the case, however. If the one panel had shown more insight into the possible sensitiveness of the exercise than had the other panels, the staff might agree that the exercise should be taken out. It should be noted, however, that loss of a specific exercise did not seriously affect the measurement of any objective, since there were a number of exercises designed for each objective.

Results

Table 3-3 shows the results of the reviews of exercises by lay panels to judge for possible offensiveness.

Of the exercises originally submitted by contractors, 1,215, or 5.8 percent, were submitted to lay panels for review. Of those exercises taken to lay panels, 24.8 percent were judged to be offensive and were recommended to be eliminated; 14.9 percent were considered possibly offensive and were revised in such a manner as to eliminate the offensiveness; 60.3 percent were judged all right the way they were. When comparing the figures for lay panel action to the *total* number of exercises written, the percentage of those deemed offensive drops to 1.4 percent and those revised to eliminate offensiveness is only .9 percent (although this figure would be slightly larger if the citizenship exercises were included in the count).

TABLE 3-3
Results of Lay Panel Reviews of Exercises[a] for Offensiveness

Subject	Approximate No. Exercises Written	Total No. Exercises Reviewed	Percent of Total No. Exercises	No. Dropped Result of Review	No. Revised Result of Review	No. Approved
Art	2468	82	3.3	1	6	75
Citizenship	600	305	50.8	8	[b]	297
Literature	2095	314	15.0	113	24	177
Mathematics	1911	0	.0	0	0	0
Music	1949	49	2.5	19	0	30
Reading	5425	176	3.2	70	38	68
Science	843	12	1.4	4	0	8
Social Studies	1776	185	10.4	79	28	78
Vocational Education	3659	90	2.5	6	84	0
Writing	283	2	.7	1	1	0
Totals	21,009	1,215		301	181	733
% of Total Number of Exercises Reviewed				24.8	14.9	60.3
% of Total Number of Exercises		5.8		1.4	.9	3.5

[a]Tabulations include figures for all exercises, including those overlapping exercises which are used at more than one age level.
[b]Citizenship exercises were reviewed and revised and re-reviewed so often that no figure is meaningful in this column. A large number of exercises underwent some revision.

A study of the table shows that some subject areas were obviously more sensitive than others. Half of the exercises written for Citizenship were submitted for lay review, and most of these were reviewed and revised a number of times. Social Studies and Literature also had a rather large proportion of exercises submitted for review (10.4 percent and 15.0 percent respectively), whereas Mathematics (0 percent), Science (1.4 percent) and Writing (.7 percent) had very few possibly offensive exercises.

Citizenship exercises were reviewed at all five conferences. Many citizenship exercises were rewritten as a result of lay panel suggestions and then reviewed at subsequent conferences. In reviewing citizenship exercises, the panelists had considerable discussion over questions concerning race, religion and sex. Some panelists felt that questions reflecting these subjects were so objectionable that their states might refuse to cooperate with the entire assessment if the questions were used. Other panelists were concerned about the effectiveness of an assessment of Citizenship if questions relating to these matters were eliminated. They contended that respect of one's fellowman is basic to good citizenship and elimination of exercises probing these areas would lead to a "pablum" type assessment. If care is taken with the wording and organization of the questions, the panelists thought some of the sensitivity could be reduced. For instance, questions involving racial attitudes were potentially offensive, but an exercise, such as

Race should not be a factor in hiring someone for a job.

Agree _____ Disagree _____

Explain, if you wish _____

was approved as it did not imply a "correct" attitude. One could respond either way and by explaining his reasons justify his response (at least to himself). Another example which was approved because of the way it was worded was

The owner of a motel or restaurant that is open to the general public should be allowed to serve only those he chooses to serve.

Agree _____ Disagree _____

Explain, if you wish _____

However, had it said

The owner of a motel or restaurant that is open to the general public should be allowed to serve only those he chooses to serve.

Do you agree with this statement? Yes_____ No_____

it might have been judged offensive because one was given no opportunity to justify his response. An exercise which was dropped by the panelists as they felt it tapped an area which is too dangerous to test at this time and which might hurt the entire assessment is the following:

Have you ever taken part in any activity intended to promote peace?

Yes _____ No _____
(If yes) Approximately when and where and what _____
(this information was to be used to validate the answer)

Art was another area which presented problems. Some of the lay people felt strongly that there should be no religious subject matter included and no nudes. Others felt that early art was mainly religious in motif and it would be wrong to omit religious subject matter. Still others felt that nudes were an integral part of art and could not legitimately be omitted. Those who opposed including nudes did so on the grounds that parents could control the types of art their children were exposed to by not taking them to art museums. However, parents would not be able to shield their children from the art displayed in assessment exercises. Lengthy discussions were held at two of the conferences. The result was a decision to include religious art and nudes, but that religious art should not be presented out of proportion to its place in history, and any nudes included should be carefully chosen. (For instance, a number of panelists objected to a painting of a mother nursing her child, and as a result, it was omitted.)

The reactions of the lay panels indicate that certain generalizations can be made regarding the types of exercises which may be considered offensive. The following classifications of subject matter indicate the types of material considered offensive:

1. Invasion of privacy	Questions in any way connected with family financing. Included would be questions dealing with how much an individual earned or received as an allowance, how income is budgeted, what contributions are made and what taxes are paid.
	Exercises dealing with a parent's relationship with his children. Included would be questions such as "When was the last

	time you had to scold or punish one of your children for not minding you?"
	Efforts to observe actual behavior with respect to health and safety rules. For instance, the inspection of homes and cars to see if they are in a safe condition, or the attempt to observe children in washrooms to see if they followed health rules pertaining to the washing of hands would be considered offensive.
	Asking teachers, counselors or employers to rate an individual's performance as compared to the performance of others was considered an invasion of privacy.
2. Minority groups	Material which might be interpreted as demeaning the Blacks (old Negro dialect) or some other minority group (reference to the size of a Jew's nose) was considered offensive.
	References to specific minority groups should be eliminated whenever possible. There should not be an over-weighting of exercises dealing with race or Blacks. More exercises about other ethnic and minority groups should be included.
3. Sex	Literary passages with sexual references or themes, such as Oedipus Rex, should be very carefully chosen as they have a great potential for offensiveness.
	Questions dealing with birth control, unless very carefully stated, should be avoided.
4. Religion	The area of religion was considered so sensitive that it could not be handled without emotional reaction.

	Reference to the "God is Dead" movement was considered too controversial to use. However, religious music could be used on tapes as long as one did not have to identify it or relate it to religion. Religious subjects could be used on art plates.
5. Human rights	Exercises dealing with human rights must be worded carefully to reduce implications which are offensive. For instance, the statement, "The number of murders increased in a neighborhood where blacks and whites lived together," can be interpreted to imply that there is more violence where blacks are. Another exercise implied that only the "poor" steal. The overabundance of exercises dealing with a person's rights was considered offensive unless more exercises were added dealing with his responsibilities in a free society.
6. Emotion-arousing terms	Any reference to sex, unwed mothers, divorce, whisky, the FBI, the President, Communism and specific organizations, such as the Klu Klux Klan and labor unions, might make an exercise offensive unless extreme care was used in the wording. The terms "protest meetings" and "demonstrations" apparently evoke emotional reactions. Preferable wording might be "public meetings on behalf of" or "in protest of."
7. Violence or cruelty	Exercises dwelling on violence or cruelty were considered offensive.
8. Words in poor taste	Passages with certain words or phrases

	were considered inappropriate for use in the assessment (e.g., a poem that includes the line, "sportive ladies leave their doors ajar," or references to dope, divorce and drunkenness for ages 9 and 13).
9. Censorship	Affective exercises in the area of censorship which indicate a restriction of a family member (e.g., "my brother") should be avoided. Questions such as "I would not allow my child to see a movie in which there is an adulterous love affair" are personal questions about one's behavior and are emotion arousing, so are felt to be an invasion of privacy. Statements about censorship in general are less offensive. "I think movies with adulterous love affairs should be banned" is an expression of an objective opinion. It does not threaten one's privacy, and hence is much less offensive.
10. Uncomplimentary comments about specific groups or individuals	Exercises which show national heroes in an uncomplimentary fashion though factually accurate are offensive. Exercises which might be interpreted as putting the police or other authorities in an unfavorable light are offensive. Negative statements about any group (e.g., "All politicians are corrupt") or attitude statements which infer that a group of people are peculiar (e.g., scientists) are offensive. Exercises about highly controversial figures, such as the late Senator Joseph McCarthy, should be neither too critical nor too favorable.

11. Inferiority of other nations	Any exercise implying the inferiority of other nations or exercises which imply the superiority of Americans to people in less well developed countries are offensive.
12. Questioning one child about another's behavior	The suitability of obtaining information from one youngster about another was strongly questioned, as it would seem to encourage "tattling" and would probably not obtain any information which couldn't be obtained in a more reliable way.
13. Darwinian theory	Questions which require the interpretation of Darwinian theory were considered objectionable. However, there was no objection to factual questions regarding the nature of Darwin's theory.
14. Civil War	In exercises about the Civil War, care must be taken to eliminate terms such as "hatred," "forced upon the South," or implications that the North was better than the South.

An example of an exercise which was dropped because the lay reviewers felt that it implied the inferiority of another nation was the following literature exercise:

Directions:

Read the paragraph below carefully and then circle the letter by the character or tale that you think the story is based on.

When the explorers arrived at the head of the river, they were surprised. They'd expected the natives to be dirty thieves, cannibals even. But they were treated with greater courtesy than they would have been in the United States. These people were ignorant, yes, but clean, generous, and highly moral.

(A) The New Breed
(B) The Noble Savage
(C) The White Man's Burden
(D) The Old Adam

Conclusion

As the National Assessment is organized to examine children and

adults in all parts of the United States, the lay people were probably overly conscientious in pointing out areas of possible offensiveness. What is offensive in one part of the country is not necessarily offensive somewhere else.

However, the wisdom of trying to locate potentially offensive material and either to eliminate it or rewrite it to alleviate the offensiveness became apparent when the assessment program got underway in March 1969. Some of the exercises which the lay people had questioned, but which they felt were of value to ask and hence did not recommend elimination, drew comments from some school districts and a few such exercises were even withheld in some states. There was one state where the problem was resolved at the state level and one district where modifications had to be made in the administration. The total number of exercises, however, which were challenged in the first assessment of 17-year-olds represented less than five percent of those used (nine out of 208).

Mail Review of Exercises by Subject Matter Specialists

As soon as an exercise was cleared as non-offensive by either the staff or a lay conference, it was mailed to a subject matter expert for further review.

Purpose

The purpose of the mail review was to have each exercise independently reviewed by a specialist who had not been involved in its construction. Hopefully the review would provide a greater breadth of view in regard to content validity and appropriate sampling of the objective. However, it became apparent that such reviews are subject to certain limitations and must be carefully planned.

Method

The reviewers were all subject matter specialists asked to review exercises in their respective subjects. Considerable care was taken in choosing these reviewers. Experts were chosen from among those people who are recognized nationally in their own fields and who had had experience in exercise writing. To obtain nominations of individuals who might be involved in this independent review of exercises, the following organizations were contacted:

1. International Reading Association
2. National Council of Teachers of English
3. National Council of Teachers of Mathematics

4. National Association of Industrial Teacher Educators
5. National Science Teachers Association
6. American Industrial Arts Association
7. American Vocational Association
8. National Art Education Association
9. National Council for the Social Studies
10. American Historical Association
11. Music Educators National Conference
12. National Association of Schools of Music

Nominations were received from all of these organizations. Invitations were sent to 41 nominees for involvement in the review of exercises. There were from three to five reviewers used for each subject.

The following procedure was used in the mail review. Originally a letter was sent to a nominee asking if he would like to be involved. The letter explained briefly the purpose and organization of the National Assessment Project, and reprints of articles about the assessment were enclosed. Upon receipt of an affirmative answer from a prospective reviewer, the staff mailed him copies of exercises in his subject area. With each exercise sent, a copy of the objective for which it was written was enclosed. As the exercises were received from the contractor, if they were judged inoffensive by the staff, they were immediately sent to a mail reviewer. At the close of each lay conference, all of the exercises which were judged to be inoffensive were mailed out to reviewers. Thus a reviewer did not receive a selection of exercises chosen for representation of all objectives for the subject. About 72 percent of the exercises originally received from the contractors were reviewed. As a few exercises had been considered offensive by the lay panels, or were not complete (for instance, in the area of Art some of the plates were missing) or were judged to be poor exercises by the staff, not all of the exercises were mailed to a reviewer, but of those mailed there was almost 100 percent return. Each exercise was reviewed by only one reviewer.

The reviewers were asked to consider each exercise from the following standpoints: (1) Does this exercise sample the objective indicated? (2) Is the answer indicated the correct answer to this exercise? (3) Does it contain any flaws that would make it ambiguous or in any way a poor exercise? and (4) What proportion of the _____-year-old people in this country do you estimate would respond correctly to this question? (The blank space was filled in with the age for which the exercise was to be reviewed.)

TABLE 3-4
Results of Mail Review of Exercises
1966-67
Number and Percent of Exercises Receiving Negative Comments

Subject	Approximate No. Exercises Received	Total No. Exercises Reviewed	*Negative Comments* Number	*Negative Comments* Percentage of Total No. Reviewed
Art	2468	2454	740	30.2
Citizenship	600	589	102	17.3
Literature	2095[a]	874[b]	298[b]	34.1
Mathematics	1911	1861	703	37.8
Music	1949	1926	255	13.2
Reading	5425[a]	2067[b]	1077[b]	52.1
Science	843	793	113	14.2
Social Studies	1776	1175	375	31.9
Vocational Education	3659	3084	648	21.0
Writing	283	244	93	38.1
Totals	21,009	15,067	4,404	29.2

[a]Overlapping exercises were counted for each age.
[b]Overlapping exercises were counted only once and not for each age.

Results

It is apparent from Table 3-4 that the percent of negative comments received from mail reviewers varied from 13.2 percent in Music to 52.1 percent in Reading. Roughly 30 percent of all of the exercises reviewed had some type of negative comment, although outright rejection of exercises was less than one percent. Typical negative comments are as follows: "ambiguous," "reading level steep," "best answer not shown," "question not clear," "rewrite stem," "not a fundamental fact" and "sketch would help." Approximately one-quarter of the exercises required some reworking as a result of these comments, and they were returned to the original contractor for such revision.

The types of comments received from mail reviewers are exemplified by the following two examples:

1. Do you think that writing has anything to do with keeping friends?

   Yes _____ No _____

The exercise was written for 9-year-olds to see whether they recognized the value of writing for social needs.

The mail reviewer commented: "I'm afraid I find this a little silly. I don't know whether the answer is yes or no, in all circumstances, and I'm afraid the question penalizes the thinking student who might see alternatives. If it were intended to produce a writing example, it might get something."

2. *Directions:* Choose the best answer.

A scientist could NOT take the temperature of
(A) Simple Simon's pie
(B) a boy or girl
(C) boiling water
(D) a hot oven
(E) a cold pond

The exercise was written for 9-year-olds to see whether or not they realized that some questions are amenable to scientific inquiry and others are not.

The mail reviewer commented: "I suppose we are to realize that Simple Simon's pie is not real. The scientist might not be real also, and thus take the temperature of the pie. I think that the entire exercise is light weight. Many kinds of questions that are not now amenable to scientific inquiry may be later. Many that were not thought to be amenable to inquiry now are. Why lead a 9-year-old to prejudge?"

Comments from the mail reviewers ranged from favorable to unfavorable and indicated areas needing improvement. Typical of the comments received are the following:

I was impressed by the creativity of certain approaches. For example, the picture or plate approach to the recognition of literary works (ranging from fairy tales to *Lord Jim*) is interesting and effective. I like your transfer of characters and situations in myth, legend and the Bible to contemporary situations (applications in advertising because the symbolism is relevant). I like the "poetry sense" tested through judgments about missing lines in poems.

You go beyond testing for memory of fact in many places in your test. Test takers are asked to apply knowledge, to sense relationships, even to make subtle judgments. In this regard, the test items sample various levels in the taxonomy of educational objectives.

I have noted some weaknesses here and there, but I am generally very pleased by the tests. You may note that in a few places I felt that your statement of objective did not give me enough information—that is, I was not adequately informed about the nature of your objective. The

cards attached hinted at objectives topically, but I was not really sure what you intended as behavioral objectives.

For many of them, the answer is obvious if a student has a naive view of a process, for example oxidation. However, if the view is sophisticated the answer is not obvious. Consequently, a number of bright students will answer questions incorrectly even though their answers are correct. I think you should look at this very carefully.

I am distressed by the rather naive views shown in the questions that have to do with laws, theories, and the like and the way in which scientists are presumed to behave. For example, the student is led to the conclusion that a scientist always begins research by making observations. Very often, this is not true at all. Scientists work on hunch, intuition, and all other sorts of beginning places.

Conclusions

The mail review was helpful in many ways. Excellent comments were received from the reviewers both for revising given exercises and for developing new ones. The review eliminated a number of exercises which had obvious flaws, and it pointed out the need for further development of easy exercises, as many of the exercises indicated by the contractors as 90 percent ones did not receive this type of rating from the reviewers.

However, it did not accomplish all that was originally hoped it would. Many of the suggestions received from the reviewers were not clear to the contractor's exercise writer who was expected to revise the exercises. Often the exercise writer had reasons of his own for writing the exercise in the manner he had, and it was difficult for him to reconcile his reasoning with the comments from the reviewers. This lack of interaction between the reviewer and writer meant that many good ideas became lost. It also became apparent that the reviewer, in order to do a thorough job, needed a better acquaintance with the assessment project than he was given. Review of exercises must be in relation to the total area and not just in relation to individual objectives. As one reviewer commented,

Apparently you sent me a portion of the total test . . . so there are at least two objectives I didn't have—II-A and III-A . . . If my comments on objectives don't make sense, it's obviously because I'm not seeing the whole picture.

In other words, the complete objectives should have been mailed to each reviewer instead of just those pertaining to the exercises being reviewed. Also, a reviewer needs to know the complete coverage of a particular objective. He cannot comment adequately on exercises when

they represent just a small proportion of those for an objective. Comments from the reviewers reflected this confusion about the total coverage of subject areas.

As the mail reviewing came to a close, the idea for conferences of subject matter reviewers developed. At such conferences, the reviewers could see all of the objectives and exercises written for a given age and thus would see the overall picture for that age. Also, there could be an interaction between the contractor's exercise writer and the reviewers, and National Assessment staff members could be present to be sure that criteria considered important to assessment exercise development were followed.

CHAPTER IV

Initial Studies

The purpose of the first of the special studies was to see if exercises designed to be "easy" (90 percent exercises) could, in fact, be answered correctly by most of a particular group of respondents. Another early study was designed to see if various proposed formats and procedures were really feasible, i.e., whether the subjects could cope with the tasks as presented. A third major problem was the feasibility of assessing individuals in an out-of-school situation. Tyler[14] gives the following pursuasive rationale for these studies:

> The need for these tryouts was clear from the initial planning of the project. Tests currently in use concentrate their exercises on the performance of "average" students. Very few if any exercises in tests now on the market are directed towards the performance of slow learners or of the more advanced. We have no way of ascertaining from comparable tests the progress being made in education by the lowest third and the highest third of school children. Hence, one of the specifications for the assessment exercises is that approximately one-third of them represent the educational performance of the lowest third of the age group, one-third represent the educational performance of the middle third and one-third of the exercises represent the educational performance of the top third. Since test constructors have not heretofore had to meet this specification it is very necessary to find out by tryouts whether the exercises do, in fact, represent this distribution.
>
> Another new feature in this project is the assessment of out-of-school youth as well as those in school, and adults, as well as children. Our plan calls for assessing the educational performance of a representative sample of nine-year-olds, thirteen-year-olds, seventeen-year-olds and adults between the ages of twenty-six and thirty-five. These ages were chosen as approximately the ages at which a majority are completing primary education, elementary education and secondary education. Furthermore, age seventeen is the oldest age at which a majority of youth is still in school. Age twenty-six is a time when the

[14]Tyler, Ralph W. *Progress report on the try-outs of the assessment exercises.* Unpublished paper, St. Paul, Minn.: Exploratory Committee on Assessing the Progress of Education, July, 1967.

great majority of adults has completed all formal schooling including professional education.

However, nearly half[15] of the seventeen-year-olds are not in school and most adults are out of school. Hence, to assess representative samples at these two ages requires the persons to take the assessment exercises in the home or other non-school situations. Heretofore, civilian tests have not often been given in out-of-school situations. Therefore, it is essential to try out procedures for assessing educational performance in out-of-school situations to see how feasible this is and how similar are the results obtained when compared with assessments in the school situation.

Such questions as the following needed to be answered:

1. Are the exercises equally distributed for different levels of achievement, including exercises for the lowest third, middle third and the upper third of each age group?
2. Are there appropriate exercises for each of the four age groups (9, 13, 17 and young adult)?
3. Is it feasible to assess individuals in an out-of-school situation?

Since the criteria for exercises represented somewhat of a departure from those usually followed in the development of standardized tests, it was only reasonable that questions should arise. As the first materials were received from contractors, it became evident that more research was needed to evaluate type of exercise and format. Writers had been encouraged to "think big" and to use new and unique formats. One immediate question was whether or not unique formats which students were not accustomed to could be easily understood. Another related question was whether or not students could handle a great variety of different formats in one assessment period. Closely related to this was another packaging[16] consideration. Should different subject areas be intermingled in one package or should all exercises in a package be of a single subject area?

Finally, and most important, were such practical concerns as:

1. Can the exercise be easily understood by the student? Does he know what he is expected to do?

[15]More recent estimates indicate there are less than 20 percent out-of-school 17-year-olds, by the specific definition finally adopted by National Assessment.

[16]"Packaging" is the process of grouping exercises together into convenient units for administration (approximately 40 minutes). The resulting exercise grouping is referred to as a "package."

2. Is the vocabulary level such that it can be understood by low achieving students?
3. Are there any problems in administration?

At this point the following three studies were done:

1. Ninety Percent Study (April-May 1967)
   This study sought to answer whether or not contractors had been able to produce enough "easy" exercises. On review, it appeared that "easy" exercises designed for the lowest achieving students might not be as "easy" as estimated. Hence the 90 Percent Study was designed and conducted.
2. Feasibility Studies — Children in School (May-June 1967)
   The Feasibility Studies of Children in School were designed to examine problems related to vocabulary, clarity, understandability and administration procedures.
3. Feasibility Studies — Adults and Out-of-School Youth (July-August 1967)

   In addition to the factors important for children in school, this study sought to determine whether questions of a cognitive nature could be successfully asked in the home or other non-school situation.

The 90 Percent Study

In the original specifications, the contractors were directed to try to produce approximately equal numbers of exercises aimed at three levels of difficulty: (1) exercises which were difficult (those which only the most able students would know), (2) exercises which were of moderate difficulty (those which the average or typical student would know) and (3) exercises which were easy (those which most students would know). This was to be done at each age level and in each of the 10 subject areas.

When the exercises were received and reviewed by the staff, the question which immediately arose was whether or not the exercises intended to be "easy" were indeed "easy."

To test this question, the 90 Percent Study was designed by the staff, and administration was carried out by the Southeastern Education Laboratory (SEL) in Atlanta, Georgia, and the Eastern Regional Institute for Education (ERIE) in Syracuse, New York.

Method

Ages 9 and 17 were selected for investigation, and it was decided that both students with high socio-economic backgrounds (high SES) and

those with low socio-economic backgrounds (low SES) should be represented in the study.[17]

In selecting exercises from each subject area to be used in the study, a random sampling procedure was used to select 54 group administered exercises written to be "easy" exercises. In areas where less than 54 "easy" exercises existed all of the exercises were used. A total of 646 group administered exercises were selected for inclusion in the study. Table 4-1 gives the number of exercises which were used by age and area.

TABLE 4-1
Number of Exercises Used in 90 Percent Study

Area	9-year-olds	17-year-olds	Total
Art	11	15	26
Literature	15	27	42
Mathematics	54	54	108
Music	10	10	20
Reading	55[a]	55[a]	110
Science	54	54	108
Social Studies	30	54	84
Vocational Education-I[b]	21	53	74
Vocational Education-II[b]	—	54	54
Writing	8	12	20
Total	258	388	646

Note—Citizenship was not included in the 90 Percent Study since there were no group administered exercises for 9-year-olds.
[a]Reading contained 55 exercises instead of 54 because an extra exercise was needed to complete a package.
[b]Vocational Education is shown as two entries because this area was divided between two contracting agencies.

Within each subject area the exercises were randomly assigned to different forms estimated to constitute 20 to 30 minutes of working time. Then two subject areas were randomly combined to produce packages containing an estimated working time of 40 to 60 minutes.

[17]Poverty level was the criteria for SES classification, i.e., subjects from below poverty level homes were classified as low SES and subjects from above poverty level homes were classified as high SES. ERIE used total family income of $4000 per annum or less to define poverty level, and SEL used a figure of $2000 per annum or less.

The packages were administered to a total of 400 9-year-olds and 416 17-year-olds by SEL in Atlanta and ERIE in Syracuse as follows:

AGE		SEL	ERIE
9	Low SES	100	100
	High SES	100	100
17	Low SES	104	104
	High SES	104	104

SEL used eight schools from six different school systems in Georgia, Florida and Alabama. ERIE used 21 schools from six different systems in New York State.

Results

Frequency distributions were prepared by age, SES and geographical location as well as for the total group for all subject areas. The distributions prepared for the overall groups of 9-year-olds and 17-year-olds are reproduced in Tables 4-2 and 4-3. These distributions show the number of exercises (all originally written to be 90 percent exercises) and the actual percent of students who could succeed on them. For example, in the area of Reading at age 9, six exercises were passed by 70 percent to 74 percent of the group and four were passed by 25 percent to 29 percent of the group.

As may be readily seen, many of the subject areas spanned almost the entire range from very difficult to very easy, yet all had been designated as 90 percent by the contractors. For 9-year-olds, Reading, Science, Social Studies and Math were examples of this. For 17-year-olds these same areas plus Literature and Vocational Education demonstrated this wide range.

Literature for 9-year-olds is particularly interesting since none of these "easy" exercises could be passed by more than 54 percent of the students. Art, Music and Vocational Education seem to come closest for the 9-year-olds in living up to their designated difficulty level, but even here there are some exercises which could be passed by less than half the students. Art and Music also seem to come closer to the intended degree of easiness among the 17-year-olds.

Tables similar to 4-2 and 4-3 were prepared for low and high SES groups. A comparison between low and high SES groups showed a range of patterns very similar to those just described for the total groups. However, the median values for low SES groups were consistently lower

TABLE 4-2
Number of Exercises by P-Value by Area for Age 9[a]

P-Value[b]	Art	Literature	Math	Music	Reading	Science	Social Studies	Voc Ed I
95% & above							3	
90 - 94	1		2		1	2	1	1
85 - 89			1	1	1	1	3	1
80 - 84			1	1	2	3		1
75 - 79	2		1			4	1	6
70 - 74			4		6	3	3	3
65 - 69	2		1	1	2	2	3	5
60 - 64			1	2	4	5	2	2
55 - 59	3		7	1	7	6	2	
50 - 54	1	2	5		3	8	2	1
45 - 49		2	4	2	6	4	3	1
40 - 44		2	3	2		6		
35 - 39	2	1	4		6	2		
30 - 34		1	3		4	2	1	
25 - 29			5		4	2		
20 - 24		1	5		2	1	1	
15 - 19		2	4		3	1		
10 - 14			1		1	2		
05 - 09		1	2		1			
4% & below					1			
TOTAL	11	12	54	10	54	54	25	21
Median	59	40	45	60	49	54	67	72

[a]The discrepancies among the number of exercises selected for the study and the number reported are due to scoring problems which were encountered with some exercises, causing them to be deleted from the study.
[b]P-Value is the percent of students who answer the question correctly.

TABLE 4-3
Number of Exercises by P-Value by Area For Age 17

P-Value	Art	Literature	Math	Music	Reading	Science	Social Studies	Voc Ed I	Voc Ed II
95% & above	5	1			2	2	1	1	1
90 - 94	1			1	2	3	5	3	4
85 - 89	4	1			3	4	3	2	1
80 - 84	1	1		3	4	3	2	3	2
75 - 79	2		1	3	8	2	2	13	7
70 - 74		3	1	1	5	2	6	5	4
65 - 69	1	2	1		6	5	2	7	3
60 - 64			2		1	8	6	3	3
55 - 59	1	1	3		7	2	6	5	10
50 - 54		2	4		3	2	4	3	1
45 - 49			9	1	4	6	3	1	4
40 - 44		2	6			2	2	2	
35 - 39		1	7		4	4	3	1	2
30 - 34		3	7		2	4	2		3
25 - 29			5		1	4	5	2	2
20 - 24		1	4		1		1		3
15 - 19		1	3		2				1
10 - 14			1			1		2	2
05 - 09									
4% & below							1		
TOTAL	15	19	54	9	55	54	54	53	53
Median	88	53	40	79	67	61	60	70	59

than the median values for the high SES groups (10 to 16 percentile points for age 9, and 12 to 23 percentile points for age 17).

As a result of this study, contractors were asked to produce additional easy exercises. Greater emphasis was subsequently placed on producing exercises which would show what nearly all students at an age level could achieve.

The important point learned from the 90 Percent Study is that even experienced exercise writers cannot "armchair" the difficulty levels of "easy" exercises. Writers, and perhaps adults in general, do not have a feel for the kinds of information which are common to almost all youngsters of a given age level. This study led to other questions such as: (1) how is the difficulty level of exercises affected by changing the distractors[18] in multiple-choice type exercises and (2) can writers successfully judge the difficulty level of difficult exercises or of exercises intended to be of average difficulty level?

A later study was designed to investigate ways in which the difficulty level of exercises might be manipulated (see Chapter VI), but no research has yet been undertaken to investigate the writers' ability to judge the difficulty level of exercises intended to be difficult or average.

This study has influenced later development by reinforcing the need to be constantly alert to problems of communication and vocabulary. That is, all exercises must be stated as simply and directly as possible so as to maximize understanding and minimize the effect of reading disability. In addition, efforts have subsequently been made to include persons in review groups who have had direct and extensive experience with low achieving pupils, either through special programs or as teachers of under-privileged children.

Feasibility Studies[19]

The first large-scale tryouts of National Assessment exercises consisted of a series of feasibility studies. A fairly large sample of exercises was tried out, prior to very much editing, with classrooms of students and with adults to find out whether there were any general problems with the exercises as a group and whether they were appropriate for the age levels for which they were designed. Contractors for in-school try-

[18]Distractors are the incorrect choices which accompany the correct answer and are intended to attract the student who does not *know* the correct answer.

[19]Adapted from: Womer, Frank B. Research toward national assessment. *Western Regional Conference on Testing Problems, Proceedings*, 1968, 34-49.

outs were the Educational Testing Service (ETS) and the American Institutes for Research (AIR). The contractor for the out-of-school tryouts was the National Opinion Research Center (NORC). These studies were conducted during the spring and summer of 1967.

Children in School

One important purpose of the studies was to look for problems that might exist when administering difficult or complex exercises to low achieving students. Presumably if low achieving students could understand the tasks, so could those of average and higher achievement status.

For the in-school sample, classes were selected to include large numbers of low achieving students at grades three, seven and 11.[20] Exercises were selected that were felt to be representative of the ones that might be difficult for students to understand. Classroom teachers administered the packages of exercises with one or two contractor's representatives present. A small number of examinees, judged to be the least able in each class, were interviewed carefully after each group administration. They were asked about the exercises and any problems they encountered. The results of the study are based primarily upon interviewer, teacher and student comments.

The major conclusion was that the exercises, as they existed in the spring of 1967, were basically usable but needed revision.

The great variety of exercise formats that had been developed proved to be both an asset and a hindrance. They were an asset in that they helped to provide much more variety in each set of exercises than a student generally is accustomed to. They were a hindrance in that each different exercise type required separate administrative directions. It was recommended that the number of types of exercise formats be reduced by using one of nine set formats whenever possible, without changing the intent of the exercise.

The mixing of subject areas (Science, Citizenship, Music, Literature, and others) within a given package seemed to be an asset rather than a hindrance. Some of the students' reactions indicated that they liked the variety of content and that they liked the types of exercises that were relatively foreign to them (e.g., listening to a musical selection and reacting to it).

The major problems with the exercises were directions that were too complex or involved and use of vocabulary that was so difficult that

[20]For convenience, grade level groups (rather than age groups) were used.

low achieving examinees had trouble understanding the task or the question. The post-test interviews were particularly productive in identifying specific words, phrases and directions that were incomprehensible to many students. The difficulty that the low achieving student has is shown by the results of the following exercise which was answered correctly by only 30 percent of the students.

Directions: Choose the best answer.
In science one is LEAST likely to study which of the following?

A. Angels
B. Plants
C. Big animals
D. Germs
E. Winds

As a result of comments made by students to the interviewer, he recommended that the word "angels" be changed to "witches" or "devils," as "angels" was too much like "angles" and caused confusion. When the change from "angels" to "witches" was further considered, the Technical Advisory Committee (TAC) pointed out that anthropologists study witches, devil-worship, and the belief in possession by devils. In addition, sociologists of religion are concerned with beliefs about devils. Rather than try to revise the exercise, the decision was made to shelve it.

Following is an example of a literature exercise designed for 9-year-olds:

Directions: This is not a test; we would like your opinion. Read the following selection carefully and then answer the question about it. Mark the *one* answer you think best.

Today is cold; the snow is falling. The only noise is a pheasant calling.

Question: Which of the following is closest to what you think about the selection?

_____(A) It belongs in a poem because it rhymes.
_____(B) It belongs in a poem because it has a regular beat.
_____(C) It belongs in a poem because it is about nature.
_____(D) It belongs in a poem because it does not move regularly.
_____(E) It belongs in a poem because it has simple words.
_____(F) It does not belong in a poem because it makes no sense.
_____(G) It does not belong in a poem because it is about snow and pheasants.

_____(H) It does not belong in a poem because it does not move regularly.
_____(I) It does not belong in a poem because it is too simple.
_____(J) It does not belong in a poem because it rhymes.

The interviewer commented, "This item has a number of words that could not be read by the pupils — *selection, pheasant, regularly, opinion. Opinion* caused the most difficulty because, even after it was pronounced, many pupils could not understand what it meant. Could we not better ask a third-grader what he thinks instead of asking him for his opinion?"[21]

A social studies exercise for 9-year-olds is presented next.

Some of these words belong to history and others are used more in geography. Write the words in the correct spaces below the word *History* or *Geography*.

dinosaurs	castles	mountains	oceans	forests
cavemen	knights	pyramids	rivers	deserts

History	*Geography*
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

The interviewer stated, "The task points up an interesting fact. The children did not know what History and Geography are because the course given in school is simply Social Studies. I can make no judgment about the task, except, of course, that it didn't work in this form."

Other kinds of information obtained are typified by the following:

1. Students do not know the meaning of "arrange in order." However, they could do the problem when told to put the smallest number in the blank next to the first word, and so on.
2. Substituting the word "song" for "melody" relieved a vocabulary problem in Music.

Contractors reported that reading an item aloud to an examinee could change lack of understanding of that exercise to understanding.

[21]Gordon, George B. *National assessment feasibility study*. Unpublished report to Exploratory Committee on Assessing the Progress of Education from Educational Testing Service, June, 1967.

Poor reading skill was a major handicap with the low ability examinees, particularly at grade three.

The feasibility of using classroom teachers for National Assessment was considered in these studies. The two contractors disagreed on this point—one said yes, it is feasible; the other said no, it is not feasible. Both agreed, however, that there were problems in getting directions to a teacher, and in getting the teacher to actually read the directions ahead of time, to follow directions exactly and to use special equipment such as tape recorders. They disagreed in their conclusion as to whether the problems could be overcome in the actual assessment.

As a result of the Feasibility Studies and in applying the general principle of "maximizing student understanding," it was decided to:

1. Hold a series of conferences during the summer of 1967, in which subject matter specialists would review each exercise for content validity and for simplicity of wording.
2. Attempt to reduce the number of unique exercise formats by adopting a standard multiple-choice format, a standard short answer format and other standard formats.
3. Move ahead with the decision to include two or three subject areas in each package for the actual assessment.
4. Investigate methods of test administrations in which both the directions and the exercises would be read aloud to examinees.
5. Investigate individual administration of National Assessment exercises.

Adults and Out-of-School Youths

The out-of-school Feasibility Study was conducted on an individual interview approach at the home of a 17-year-old or a young adult (ages 26-35). The youth or young adult was asked to participate in an educational study by answering some questions. The impetus for this study was the fear that adults, in an interview situation, might not be willing to answer a series of primarily cognitive questions. It was felt that while adults generally are happy to express their opinion about something, exposing themselves to a potential wrong response might lessen the percentage of cooperation.

Three packages were assembled so that all types of exercises, all 10 subject areas and all levels of difficulty were included. Field work was done in high-income suburbs, low-income central city areas and in a

middle-sized city and the rural areas near it. Regular survey interviewers were used.

Comments from the interviewers about their experiences administering the exercises were helpful in pointing out need for revisions or changes in administration.

One interviewer made the following comment on music exercises administered to adults:

> [The music exercises] asked if respondents recognized a piece of music, and, if so, whether they knew its name and composer. Some respondents evidently construed "recognize" as "know the name of" and said that they had heard a piece but could not identify it. The question should be revised as follows:
>
> Have you ever heard this piece of music? Yes No
>
> IF YES: Who wrote it? _____I don't know (remember)
>
> What is its name? _____I don't know (remember)
>
> There were very few right answers to the two sub-questions of the original exercise.
>
> In general, respondents seemed to enjoy the questions relating to music tapes. Even so, the interviewers thought that the selections were too long and noted that some respondents began writing before a tape had ended.[22]

The major conclusion of this study was that it is feasible to administer exercises by a personal interview carried out in a respondent's home. None of the exercises was impossible to administer and there was little reluctance to begin the task or to carry it through.

Additional conclusions were:

1. Much editing of exercises is needed in tone, clarity of instructions and vocabulary.
2. Respondents are not troubled by varying content areas. Grouping exercises by item format was suggested.
3. Having respondents mark multiple-choice questions themselves was satisfactory.
4. Open-end exercises should be asked orally. When answers are cursory, interviewers should probe for a more complete response.

[22]Johnstone, John W. C. and Spaeth, Joe L. *Administration of test exercises in the home.* Unpublished report to Exploratory Committee on Assessing the Progress of Education from National Opinion Research Center, Sept., 1967.

5. Packages of an hour's length or longer are feasible.

Results from the out-of-school Feasibility Study were encouraging enough to move ahead with the original plans to assess out-of-school 17-year-olds and adults. In addition, the decision to seek further editing of exercises was reinforced. The question of the potential use of mobile vans was raised. That suggestion is still being explored because it might improve measurement for exercises requiring equipment that could not easily be taken into a home for an interview.

The suggestion of varying exercise difficulty level was incorporated into packaging criteria.

Summary

The first studies to be made which influenced exercise development were the 90 Percent Study and the Feasibility Studies.

Writers are not ordinarily required to produce large quantities of "easy" exercises. The purpose of the 90 Percent Study was to determine whether "easy" exercises were easy. The results of the study indicated that writers cannot reliably judge the difficulty level of exercises intended to be easy. As a result of the 90 Percent Study contractors were asked to try again to produce additional easy exercises and greater emphasis was subsequently placed on producing exercises which nearly everyone could pass.

The purpose of the Feasibility Studies was to determine whether there were any general problems with the exercises relating to understandability or administration and whether they were appropriate for the age groups for which they were designed. Major problems were discovered in the complexity of the directions and the difficulty of vocabulary.

As a result, a series of review conferences was planned for the purpose of reviewing, editing and simplifying vocabulary and formats.

CHAPTER V

Subject Matter Reviews

By late spring of 1967, the decision was made to undertake large scale reviews of exercises at subject matter conferences. These conferences were for the express purpose of reviewing each exercise for content validity and for simplicity of wording and to provide an opportunity for interaction between reviewers and exercise writers. It would also give the subject reviewers a chance to see the complete set of objectives and exercises.

First Subject Matter Review Conferences

As the operational plans still called for the completion of all preliminary work on exercises by February or March of 1968, these review conferences were organized for the summer months of 1967 to leave time for additional work by the contractors before the end of the year. The number of reviewers per conference varied from three to 16. One or more of the contractor's exercise writers were present to answer questions about scoring, intent of exercise and so forth, and National Assessment staff members were also present.

There was some flexibility in the organization of the conferences. At some, the reviewers worked together as a group and went over the complete set of exercises; at others, the reviewers divided into smaller groups, each doing a portion of the exercises, and then met together later to discuss the overall picture. At each conference the exercises were rated as follows:

Key for Evaluation of Exercises

Symbol	*Meaning*
++	concept: important, appropriate for age intended exercise: written well; vocabulary simply (needs no editing)
+	concept: important, appropriate for age exercise: needs minor revisions

0	concept:	important, appropriate for age
	exercise:	needs virtually complete rewriting
-	concept:	less important, or inappropriate for age (perhaps appropriate for later age)
--	concept:	trivial (not salvable)

These ratings were made on the basis of content validity, clarity and appropriateness of content and language used for each age level. If time permitted after the review was completed, the reviewers tried to edit those exercises which needed it.

The following example from the area of Social Studies gives an indication of the type of comment made at the conferences. This particular exercise was developed to assess a 9-year-old's ability to "understand some of the major characteristics of the geographic (spatial) distribution of man and his activities, and some of the major characteristics of man's interaction with the physical environment." It was written as an exercise which 50 percent of the 9-year-olds should be able to answer. The exercise read as follows:

> *Directions:* The land about us affects the way we live. Here is a short story that tells about a boy, Wemba. Read the story and choose the answer that gives the best reason for Wemba's way of life.
>
> Wemba lives in a dark forest near the equator. He has never seen snow or ice, but he has often played in the rain. Wemba's playground has tall trees and tangled vines that make a good home for chattering monkeys and colorful birds.
>
> Wemba cannot do his homework at night because
>
> (A) he watches television at night
> (B) fire is his only light
> (C) his family cannot help him
> (D) the dark makes him afraid

The reviewers wanted the exercise withdrawn. They felt it was irrelevant to the objective and that the choices were not related to the story, and hence there was no correct answer.

Another example of an exercise receiving comment was the following one written for 13-year-olds to assess knowledge about reproduction and biogenesis:

> Small worms are seen crawling out of the mud along the river banks. Where did these worms come from?

(A) If it is warm and just right they can be formed from the mud.
(B) Worms of this kind come from the air.
(C) These worms come from the bodies of animals where they were formed.
(D) The worms can come only from other worms, never from anything else.
(E) God is constantly creating these worms out of nothing.

"Keep God out of these questions," was one comment. Otherwise, the reviewers wanted the wordage cut down and suggested the following choices:

(A) from the mud
(B) from the air
(C) from the bodies of other animals
(D) from other worms like themselves
(E) from non-living materials

Still another example is a science exercise written for 17-year-olds to measure fundamental facts and principles of Science — evolution:

At present, evidence for organic evolution is LEAST likely to be provided by a study of

(A) fossil records
(B) computer science
(C) geographical distribution of organisms
(D) embryology
(E) comparative anatomy or homologous structures

The reviewers suggested that a more plausible choice than (B) computer science, should be offered. They reasoned that computer science in and of itself did not *provide* evidence. However, a computer could be *used* to provide evidence and was therefore not necessarily a clear choice of the LEAST likely.

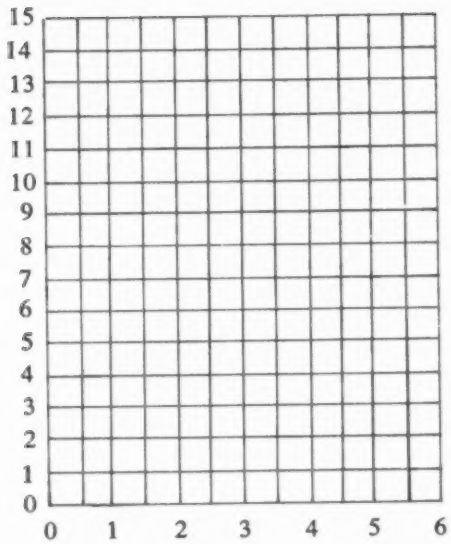
An exercise which was approved by the reviewers is the following one for 17-year-olds using science apparatus and attempting to find out whether the student understands the role of scientific laws.

Hooke's law says that there is a relationship between the distance an elastic material stretches and the force applied to the material. Find out how well the elastic cord . . . follows Hooke's law. Place different numbers of weights in the pan and measure the length of the elastic material.

Show your results by making a graph.

Graph I

Elastic Cord



These summer conferences had pointed up many problems. For some subjects the reviewers considered the materials had just scratched the surface, that they were "many times too textbookish," and that they required only rote memory. In other subjects the exercises did not show enough imagination; frequently the language was pedantic, not precise and sometimes even wrong. One reviewer commented, "In an effort to be precise, item stems[23] are long, involved and difficult to read." Most reviewers agreed that the exercises were too middle class and much of the

[23]The stem of an exercise is that portion of the exercise which states the problem or asks the question. For example, in this exercise,

Who was the first President of the United States?
☐ John Adams
☐ Thomas Jefferson
☐ Abraham Lincoln
☐ George Washington

"Who was the first President of the United States" is the stem.

vocabulary was too difficult. At all of the conferences it was obvious that more work was needed on the exercises to provide better coverage of the objectives and to meet the criticisms of the reviewers. Typical of the comments made by the reviewers are these made at one conference:

> It was apparent that the item writers were out of touch with reality from the standpoint of language in the stems and distractors. Kids will simply get all tangled in the words unless they are made simpler. *Negatively stated* questions, or stems, are not going to clear the language problem up. If possible, get rid of all of them! I will only admit to agreeing partly . . . that a child should be able to choose the WRONG item out of a list instead of the RIGHT one. There is no sense in teaching children how to take a test in this National Assessment. Kids are so conditioned to looking for the RIGHT item out of a list, that very strong bias will be introduced doing things any other way.
>
> Whenever possible items depicting real life situations might be preferable to the more "textbookish" type, especially in view of the fact that many of the adults will have been out of school for a period of several years.

In all areas the contractors were asked to rework the exercises and to produce more exercises for certain objectives. Knowing, then, that contractors would be producing new materials, plans were begun for a second round of subject matter review conferences.

Second Subject Matter Review Conferences

The summer reviews strongly pointed up the need to involve people who are working directly with low-achieving students, as such persons could give needed direction on appropriateness of vocabulary and should be able to offer suggestions for non-middle-class-oriented materials. In general, the second conferences held the following winter were organized by subject area into four panels, one for each age level. The members of each panel were chosen to include at least one person with experience in working with low socio-educational children, one person with a background in measurement in the subject and two or more who were subject experts at either the appropriate teaching level or at colleges of education with specialization in the specific subject area. Each panel had a staff member who acted as a recorder, and contractor's representatives were available to consult with the panels as requested. The emphasis at the conferences again was to examine the content validity of each exercise, the appropriateness of the content for the age, the overall coverage of each objective, the appropriateness of the esimate of difficulty and the

understandability of the vocabulary. By this time, the principle of "do everything possible to maximize student understanding of the tasks he is asked to perform," was fully ingrained, and clarity of vocabulary and directions was stressed.

Two further considerations were undertaken at the conferences. The decision had just been made to restrict "clustering" (see Chapter II). Of those exercises which were clustered, the reviewers were asked to consider which could be declustered meaningfully. The other problem arising at this time was that of "directionality" (see Chapter II). In reviewing exercises each one was to be judged on the basis of whether or not there was a right or wrong answer or a definitely preferred direction to the answer. As pointed out earlier, most of the difficulty with directionality is on attitude exercises. One must be sure there is a preferred direction to the attitude being assessed or else it is impossible to show whether or not there has been "progress" in the educational attainment of that attitude over time. With exercises assessing knowledge, one assumes that the attainment of the knowledge is the preferred direction. This time the rating of each exercise was on a three-point scale: (1) all right as it is, (2) needs editing and (3) no good. The reviewers used the following guidelines in judging and rating each exercise:

Criteria for Judging Exercises

1. Does the exercise have content validity? Is it assessing something important and desirable to know?
2. Does the exercise measure the objective for which it was intended?
3. Are the objectives well covered?
4. Is the exercise clearly stated so that the individual will understand what he is to do?
5. Is the exercise appropriately designated for group or individual administration? If an exercise needs an individual interviewer to clarify the task or to probe for certain responses, the exercise should be designated for individual administration.
6. Is the estimated difficulty level appropriate? Each exercise is estimated to be at or near one of three levels of difficulty: (1) easy (90 percent), (2) medium (50 percent) or (3) difficult (10 percent). Does your judgment agree with that of the exercise writer?
7. Does the exercise have directionality? That is, can it be scored right or wrong or on some kind of scale? Many attitudinal type exercises have problems meeting this criterion.

In general, the reviewers at the second review conferences were favorably impressed by the progress which had been made since the first conferences. Comments ranged from very favorable for Writing to general uneasiness with Art. The writing reviewers were quite positive in their reaction to the exercises and indicated that the writing package was well prepared. The art reviewers felt that most of the exercises for two objectives needed to be reworked and the reproductions of works of art should be truer to the original.

At the close of the conferences, five of the subject areas appeared to need only minor refinement before use in the assessment. These were Writing, Citizenship, Literature, Science and Social Studies. The other five subject areas still needed extensive work.

Special Problems

Problems calling for special attention or decisions arose in all of the areas. In the area of Citizenship, the concern for communicating with low socio-educational students led to special research in the summer of 1968. The contractor asked groups of low socio-educational students to review citizenship exercises for wording and comprehension. A total of 11 students were hired from a Black Community Relations Association, a summer job service for Black youth in a low socio-educational area. The 11 represented ages 9, 13 and 17, and included both boys and girls.

Two members of the contractor's staff supervised each session. Each student was given all of the exercises. Students were first asked to read the exercise to themselves silently and to circle any words they did not know. The supervisors noted these words. The circled words were then read aloud to the students and words which the students still did not understand were noted. As some students tended to circle all long words, it was necessary to read the words aloud to find out exactly which ones were not understood. After unknown words were explained to the students, they were asked to give possible answers to the exercise. Answers were noted only if they revealed a misunderstanding of what the exercise asked. Any questions the students asked were also noted. Before leaving any exercise, students were asked how it might be reworded so that "other students could understand it better." At the 9-year-old level, the three students were such poor readers that each exercise had to be read aloud before they were able to answer. As a result of this careful and detailed study of the vocabulary of each exercise, the contractor's exercise writers were able to edit the exercises and make them more understandable and

readable to the low socio-educational student. Understandability is essential, for if a student does not understand an exercise, he is not able to demonstrate whether or not he possesses a certain skill or bit of information.

The science reviewers felt that too many of the exercises asked for isolated facts and did not call for an understanding of scientific principles. Consequently, some of the reviewers were asked to write exercises. When the exercises were written, they were reviewed by mail and then given to the contractor's exercise writer for comment and proper formatting.

The social studies reviewers were also concerned over the isolated, often trivial facts which were being assessed under Objective IV: Has knowledge relevant to the major ideas and concerns of social scientists. They felt that there were too few exercises on understanding relationships and that a need existed to add exercises on knowledge of concepts and relationships. They also wanted to be sure that the exercises represented a balance of major ideas and concerns of social scientists. Three of the reviewers for age 13 and two for age 17 offered to write exercises of the type they felt were needed. As their exercises were written after the subject review conference had been held, the exercises were reviewed by mail by four subject matter reviewers. Following this special mail review, the exercises were sent to the contractor to be added to the exercise pool.

The mathematics reviewers at the conference held in the summer of 1967 felt that the language used in many of the exercises was too formal and needed to be stated more simply. The contractor argued that the wording must be precise in order to be mathematically correct. The question was then asked, "Does it make any difference how an exercise is worded as far as performance is concerned?" A study was undertaken by the staff to compare formal versus informal everyday language in the mathematics exercises. A full report of this study is given in Chapter VI. Briefly, the results showed that the wording of the mathematics exercises in that study made little difference in the percentage of correct responses made.

At the first art conference, two basic questions about the art exercises were raised which indicated the need for further consideration. The first question pertained to the quality of reproduction of the work of art necessary to elicit a valid response comparable to that elicited by the original art work. Should color slides, museum color postcards or postcard-sized black-and-white plates be used? The second question was in regard to actual performance tasks. Is performance on a two-dimensional art work comparable to that on a three-dimensional one? The difficulty of

shipping three-dimensional works to a central location for scoring is obvious, and if the two types of art works are comparable from the point of view of judging art ability, there would be no need to include exercises calling for three-dimensional art tasks. The first question was resolved by deciding to use color postcards and 11-inch x 14-inch prints. The problem of type of performance exercise is still unresolved and will be taken into consideration in the development of new exercises calling for performance of art tasks.

The value of the subject matter review conferences is obvious from these examples of special problems which were pointed up by the conferees. In some subject areas more exercises were needed, usually of a different type which called for the use of new exercise writers. In almost all areas more attention needed to be given to the vocabulary used. The reviewers were able to indicate where the vocabulary was too difficult and, in many cases, actually reword the exercises to simplify the vocabulary. In other cases, the contractor's writer was made aware of the need for re-editing and simplification. In some areas problems arose which needed research to find the appropriate solutions, and in other areas, the problems could be resolved by the staff or their advisors.

CHAPTER VI

Other Studies

While many additional studies were conducted in the preparation phases of National Assessment, three stand out as being particularly important in helping to guide exercise development. They are: (1) the Mathematics Study, (2) the Choices Study and (3) the Final Field Tryouts.

The Mathematics Study[24]

A major concern of reviewers of the mathematics exercises was their formal wording. Mathematicians insisted that wording be precise and accurate; mathematics educators were more inclined to opt for less precise language that is more easily understood by students. In order to investigate this concern, the National Assessment staff designed and conducted a study comparing a set of exercises worded formally with the same exercises worded in simpler language. The main question which this study attempted to answer was, "Does it make any difference how an exercise is worded as far as performance is concerned?"

As the design for the study was being considered, two additional questions were raised:

1. What is the effect of including an "I don't know" choice for multiple-choice exercises?
2. How does performance on multiple-choice mathematics exercises compare with performance on their open-end counterparts?

Both of these questions relate to measurement problems concerned with guessing. The use of open-end exercises had been encouraged since if one wants to find out what a person knows, one asks him to recall, not to recognize and select the correct answer from a list of alternatives. However, the multiple-choice format is being used to a considerable ex-

[24]Adapted from: Womer, Frank B. Research toward national assessment. *Western Regional Conference on Testing Problems, Proceedings*, 1968, 34-49, and Knapp, Thomas R. *The mathematics study*. Unpublished report, Exploratory Committee on Assessing the Progress of Education, 1968.

tent in National Assessment exercises. When a question is asked in multiple-choice form there is always an unknown element of success produced through guessing. Given 100 four-choice exercises a person who randomly selects answers without any knowledge of the topic will get 25 correct on the average. On any given exercise the chances are one in four that he will succeed. Different "corrections for guessing" have been proposed in measurement literature but are generally criticized on the grounds that they assume the individual has "no knowledge" about the topic and do not reflect partial information or misinformation held by the individual. Since both multiple-choice and open-end exercises have been produced for the assessment, questions related to difficulty level of the two forms were raised, and ways to reduce guessing in multiple-choice exercises became a concern.

Subjects

The subjects chosen for this study consisted of 110 high school juniors and seniors and 60 individuals who had dropped out of school,[25] distributed as follows:

School		
1	12th grade Advanced Mathematics (one class)	23
2	11th grade Regular Mathematics (one class)	27
3	11th grade English (one class)	27
4	11th and 12th grade Practical Mathematics (two classes)	33
5	Dropouts (sample of registrants) at a Youth Opportunity Center	60
	Total	170

Exercises

A sample of 32 exercises was selected. Eight of these exercises had been written as easy (90 percent) arithmetic (computational) exercises for 9-year-olds and were put at the beginning of the booklet. As most of the 32 exercises would probably be very difficult for the "dropout" group, a few easy exercises at the beginning might prevent them from getting discouraged altogether.

The other 24 exercises had been written to span a range of difficulty for 17-year-olds (roughly one-third 90 percent's, one-third 50 percent's and one-third 10 percent's) and covered a variety of topics (sets, algebraic manipulations, trigonometric concepts, and others).

[25]This sample is weighted more heavily with "dropouts" than either the actual assessment sample or the population.

Three additional versions of each exercise were constructed by the National Assessment staff. One of these versions involved the simplification of the technical language following the suggestions of reviewers. The second and third versions were formed from these two basic versions (the exercise as originally written and the simplified version) by creating an open-end parallel exercise if the given exercise was multiple-choice, or by creating a multiple-choice exercise if the given exercise was open-end.

These exercises were then grouped into four packages or booklets so that each contained 16 contractor prepared exercises and 16 simplified exercises and had the following characteristics:

Package	*Characteristics*
A	All multiple-choice with "I don't know" distractor added.
B	All multiple-choice without "I don't know" distractor.
C	All open-end with instructions to write in "I don't know" if this was the case.
D	All open-end with no instructions regarding "I don't know."

Procedure

Within each of the groups of individuals listed above, the packages were randomly assigned by cyclic arrangement[26] before distribution, so that there was an approximately equal number of subjects for each package within each group.

A National Assessment staff member administered the booklets to two of the classes. The regular teachers administered the booklets to the remaining regular classes. The "dropouts" at the Youth Opportunity Center were assessed individually by the staff at that center who were directed to read any words or phrases those subjects were not able to read, since reading comprehension was not being measured.

Each student was asked to go through the booklet twice, the first time recording his responses just as he would on a regular test and the second time circling and/or commenting upon any word, phrase, symbol

[26]In a cyclic arrangement, a random starting point is first determined. In this case, one of the four packages, A, B, C or D was randomly chosen. For example, if C was chosen, then the booklets were physically arranged in order CDABCDAB . . . , and distributed to the students in each class.

or other item which he did not understand. The student was also asked to suggest any rewording of the exercises or directions which would simplify understanding of the tasks by future students.

Findings

Three general findings emerged from this study:

1. The wording of most exercises made little difference. Easy exercises were easy no matter what wording was used; difficult exercises were difficult regardless of wording.

2. Guessing seemed to be reduced when examinees were given the option of responding "I don't know." In 23 of 32 exercises there was closer correspondence[27] between the open-end version and multiple-choice version using the "I don't know" option than there was between the versions not using the option.

3. Open-end exercises were, in general, more difficult than their multiple-choice counterparts. There were small differences between very easy and very difficult exercises but those between these two extremes showed large differences. This suggests that when a student has developed an arithmetic or mathematical skill or when he does not have the skills, then exercise format is immaterial. In the middle ranges of skill development the multiple-choice mathematics exercise is considerably easier than the open-end exercise.

The "soft data" or comments made by the students contributed additional interesting information. These are just a few of the many examples.

> Test for 8th graders or Army Inductees. (male student in Advanced Math class who answered correctly 30 of 32 exercises)

> None of these solution sets are correct. That's a lousy trick to pull because when you're taking a test you're already pretty nervous when you start, and not being able to find an answer doesn't help. (female student in Regular Math class)

> It's amazing how much algebra you can forget in two years. (female student in English class)

> I think the test could have been used for younger girls and boys of 15 years and harder test for older ones. But I think this one could be

[27] i.e., the difficulty levels were more nearly alike.

used by the ones that don't understand anything about math. (female "dropout" who answered correctly 9 of 32)

This test is much too hard for a High School Drop-out—I only went thru 9th and 7 mo. of 10—most of the problems I can't even begin to understand. I learned some equasions (sic) and graphs—but they were much simpler than these—

This study has influenced the development of National Assessment exercises in a number of ways. It has reinforced the desirability of emphasizing simplicity of wording and understandability of the task. If precise mathematical statements do not have any apparent advantage over less formal presentations, it is probably wise to heed the advice of mathematics educators.

The use of the "I don't know" option has become a standard part of all multiple-choice exercises. The open-end exercise continues to be regarded as a form worthy of developing, provided an acceptable scoring rationale and key can be predetermined.

The Choices Study[28]

The Choices Study was undertaken to investigate further the relationship between the multiple-choice type format and the open-end format. Another important dimension of this study was whether easier exercises could be produced by changing the distractors. Distractors are generally developed using one of two methods. The exercise may be tried out using an open-end version and the incorrect responses noted and tabulated or categorized. The most frequently used incorrect responses are then used as distractors in the multiple-choice form. The other approach is to "armchair" or anticipate what distractors might attract a student who has only partial information or misinformation about the topic being assessed. Only the latter approach has been used to date in developing exercises for National Assessment.

The process of "creating easier exercises" by changing the distractors if carried to the extreme implies that such manipulation may degenerate to the practice of selecting such ridiculous distractors that virtually anyone could answer any question. It is also possible to increase the difficulty by the choice of distractors.

[28]Adapted from: Womer, Frank B. Research toward national assessment. *Western Regional Conference on Testing Problems, Proceedings*, 1968, 34-49, and Knapp, Thomas R. *The choices study*. Unpublished report, Exploratory Committee on Assessing the Progress of Education, 1968.

Additional information was needed as to how the difficulty level of multiple-choice exercises could be affected by the selection of specific distractors.

To investigate this question, the staff developed a study that compared three versions of a multiple-choice exercise with the same exercise in open-end form.

If the differences among the difficulty level for exercises having the same stem and correct answer, but different distractors, are small and close to the difficulty level for the open-end version, it does not matter which way the exercise is phrased for the assessment. If the differences are large and/or if the multiple-choice difficulty levels are quite different from the open-end version, it obviously does matter.

Subjects

The majority of the subjects were eighth-grade pupils of average and below-average ability at a junior high school in Minnesota. Additional subjects at a junior high school in California, also of average and below-average ability, were also included, both to enlarge the overall sample size and to serve as a partial check on any local bias that might appear in the data. The subjects were distributed as follows:

Minnesota	206 pupils	(8 classes)
California	71 pupils	(3 classes)
Total N =	277	

Exercises

Social Studies was the area selected for investigation. A sample of 21 multiple-choice exercises was chosen; nine of these exercises involved the identification of persons, four involved the identification of countries, and three involved the identification of wars. The remaining five were miscellaneous. These exercises were all factual knowledge exercises which lend themselves readily to the construction of alternate versions with different distractors. Exercises calling for comprehension and analysis, for example, are not as adaptable to this purpose. The results obtained in this study may not generalize to such exercises, nor to other subject areas.

Four versions of each exercise were administered: the contractor's version and three new versions which were constructed. Of the constructed versions one comprised a set of distractors thought to be more homogeneous than the contractor's distractors (the expectation being that the exercise would be more difficult); one consisted of distractors

which were along a different dimension from either the contractor's version or the more homogeneous versions; and the other one was the open-end version. An "I don't know" choice was added to each multiple-choice exercise.

To illustrate, using one of the exercises actually included in the study:

Contractor's version (distractors all Latin American countries)
What is the name of the country where the Inca Indians lived?

- ☐ Brazil
- ☐ Mexico
- ☐ Panama
- ☐ Peru

- ☐ I don't know.

Homogenous version (distractors all neighboring South American countries)
(same stem)

- ☐ Boliva
- ☐ Chile
- ☐ Ecuador
- ☐ Peru

- ☐ I don't know.

Other-dimension version (distractors are distant countries)
(same stem)

- ☐ India
- ☐ New Zealand
- ☐ Peru
- ☐ South Africa

- ☐ I don't know.

Open-end version

What is the name of the country where the Inca Indians lived?

The exercises were arranged in random order and four forms were prepared. Three of these were multiple-choice forms with the various versions counter-balanced within form.[29] The fourth form was entirely

[29]Form 1 contained the contractor's version of exercise number 1, the "homogeneous" version of number 2, and "other-dimension" version of number 3, etc.; Form 2 had homogeneous number 1, other-dimension number 2, contractor's number 3, etc.

open-end. The open-end form contained "fill in the blank" directions, telling them to write in "I don't know" if they did not know the answer.

Procedure

Within each classroom the forms were randomly assigned by cyclic arrangement before distribution, resulting in an approximately equal number of subjects answering each form. There was no time limit, but all pupils finished well within the typical class period.

Results

The general results were as follows:

1. Six of the exercises were very easy in any version. For two others there was little difference in the difficulty level for the multiple-choice versions but a substantial difference between these and the difficulty level for the open-end version. For the other 13 exercises the difficulty levels had a considerable range (Table 6-1).

2. The open-end version was generally the most difficult version. This is to be expected since the task required recall rather than recognition.

3. The percentage of "I don't know" responses varied from exercise to exercise and from version to version of the same exercise (Table 6-2). One very interesting result was that, of the 208 pupils who took 21 multiple-choice exercises each, there were only four exercises omitted, that is, ones left unanswered by the pupils; of the 69 pupils who took 21 open-end exercises each, there were 198 exercises omitted (the percentage of omits per open-end exercise ranged from zero to 33). There was a greater tendency for pupils to make use of the "I don't know" option with multiple-choice exercises than with the open-end version.

The Choices Study has supported two previous findings of the Mathematics Study. The open-end version tends to be more difficult than the multiple-choice form except for very easy exercises. (While this finding is nothing new to the field of educational measurement, it is one which needs to be studied in different contexts.) The "I don't know" option does tend to be used if the opportunity is provided in a multiple-choice exercise, more so than in an open-end exercise where instructions are given to write in "I don't know." This finding lends additional credence to the fact that guessing is reduced by using the "I don't know" option with multiple-choice exercises.

The most important implication of this study is that the difficulty level of an exercise can be manipulated by the choice of distractors or the use

TABLE 6-1
Difficulty Levels
Obtained in The Choices Study

Exercise	Version			
	Contractor	Homogeneous	Other-Dimension	Open-End
1	92%	81%	87%	52%
2	99	100	100	99
3	72	76	85	62
4	99	99	100	94
5	75	46	70	16
6	39	29	45	49
7	73	61	60	59
8	25	25	20	3
9	39	24	25	6
10	21	28	53	6
11	63	69	86	68
12	97	88	100	96
13	99	99	97	96
14	37	14	38	10
15	90	90	94	87
16	56	54	71	19
17	51	48	62	20
18	72	69	68	58
19	23	10	40	6
20	28	38	29	29
21	94	94	92	86

Note.—The difficulty level is expressed as the percent of pupils giving correct response (i.e., 92 percent of the pupils responded correctly to the contractor's version of exercise number 1).

of the open-end format. Because of the difficulty encountered in attempting to create "easy" (90 percent) exercises, the temptation is ever present to create exercises with ridiculous distractors.

The very serious question is then raised as to whether reasonable and logical distractors can be "armchaired" or whether the practice of administering a question in open-end format to obtain logical distractors is a better procedure. The study also indicates a need to define better just where the multiple-choice form is to be preferred over the open-end

TABLE 6-2
Percentage of "I Don't Know" Resporses

Exercise	Version				Omits for Open-End Version
	Contractor	Homogeneous	Other-Dimension	Open-End	
1	6%	12%	9%	25%	21%
2	0	0	0	0	0
3	4	13	4	4	10
4	0	0	0	0	3
5	16	18	19	29	25
6	23	18	38	17	14
7	3	4	4	3	4
8	49	54	68	52	32
9	17	26	35	32	25
10	15	16	34	12	7
11	9	7	7	6	6
12	0	0	0	1	1
13	1	0	1	3	0
14	43	24	48	33	23
15	4	4	1	3	3
16	4	12	4	10	14
17	18	20	19	20	25
18	12	12	15	4	13
19	34	33	47	29	33
20	28	38	35	12	20
21	0	0	4	3	6

form. Neither of these questions has been adequately investigated at this stage of development in National Assessment.

Final Field Tryouts Before Assessment

In the spring of 1968, five of the 10 subject areas were ready for the final tryouts which preceded the assessment: Citizenship, Literature, Science, Social Studies and Writing. Earlier feasibility and research studies and the lay and subject matter reviews had led to much revision. It was necessary to determine what problems still remained with the revised exercises.

The final tryout of exercises was as follows:

Ages 9 and 13:[30]

1. American Institutes for Research (AIR) tried out the individually administered exercises with an N of about six for each exercise. Feasibility type data only was sought. That is, emphasis was on determining the understandability of the task and on adjusting the vocabulary to an easier level if the need became evident.

2. Educational Testing Service (ETS) administered exercises in groups with an N of about 60 for each exercise (30 high SES, 30 low SES).[31] A maximum of 30 group exercises per objective per age were tried out. Where there were more than 30 exercises per objective, a representative subset was selected. The data obtained included information on difficulty levels and general information regarding procedures.

Age 17:[32]

1. National Opinion Research Center (NORC) individually administered exercises in the home with an N of about six for each exercise.

2. ETS tried out the group exercises following the same plan as for ages 9 and 13 above.

Adult:

1. NORC tried out the exercises which were unique enough to require feasibility information, with an N of about six for each exercise. These were administered in the home using standard interview techniques.

Group Administered Exercises

The group administered exercises were assembled into 107 packages by the ETS Test Development Division staff. The packages were intended to simulate as nearly as possible the actual assessment plans. That is, each package contained a mixture of subject matter areas, contained both objective and open-end exercises and was designed to take approxi-

[30]For convenience, grade level groups rather than age groups were used in the administration. Hence, third graders were used to represent 9-year-olds (the most common grade in which 9-year-olds are found); seventh graders were used to represent 13-year-olds.

[31]For the purpose of this study high SES was defined as an estimated family income of $10,000 per year for at least 50 percent of the group and low SES was defined as at least 50 percent of the group eligible for Title I funded projects. Title I projects under the Elementary and Secondary Education Act provide special monies for educationally underprivileged school districts.

[32]Eleventh graders were used to represent 17-year-olds.

mately 40 minutes administration time. By this time the decision had been reached to tape the administration procedures as well as the exercises in the actual assessment[33] and tapes were produced to accompany each package. The tapes included both the directions for administration and the reading of each exercise.

The administration of the packages to third, seventh and 11th graders was carried out through six ETS field offices. Each package was administered to two groups, one identified as predominantly high socio-economic status and the other as predominantly low socio-economic status according to criteria previously stated. Each package was administered to a minimum of 50 individuals with not less than 20 in either sub-group. Both boys and girls were included in the groups selected.

During the administration the classroom teacher and an ETS representative each completed an observation form. The forms were designed to:

1. Obtain judgments on the reaction of the students to each question. Five specific behaviors were to be catalogued (inattention or boredom, "cutting up" or other misbehavior, inappropriate laughter, failure to follow directions and inability to cope with task). The form also allowed for "other" observations and comments about the questions.

2. Obtain judgments on overall interest level of the students, the test difficulty (both in terms of content and presentation) and the effectiveness of the taped presentation. (ETS observation forms also allowed for an evaluation of school cooperation.)

Although many persons involved in the tryouts were initially skeptical about the use of tapes, particularly at the seventh and 11th grade levels, the reactions of those people using them were generally positive. Both the teacher and the field coordinators administering the packages offered far more positive than negative statements regarding the tapes, although there were some problems noted in pacing. That is, the timing tended to be too liberal for essay tasks for low SES youngsters (they wrote all they could in a very short period of time) and too liberal for multiple-choice exercises for high SES students. Specific comments made by field co-

[33]The question of administering the assessment by tape first arose because of the ongoing concern that it was necessary to maximize student understanding and minimize demands upon reading ability. The feasibility of using a taped administration was researched in a Methods of Presentation study done by AIR. On the basis of the findings of that study a firm decision was made to tape administer all group packages.

ordinators are presented in a separate section which follows (Selected Reactions Offered by Field Coordinators).

Scoring. Exercises were placed into one of three classifications for scoring:

1. Professional. These exercises were essay-type, scored using standard ETS procedures. The scorers were college graduates and specialists in the subject area and/or at the grade level being scored.
2. Semi-professional. These exercises were generally open-end, required a written short answer response and were scored using standard ETS procedures for the reading of exercises to be scored by non-professional readers. The readers had some post-secondary education.
3. Clerical. These exercises were typically multiple-choice and were scored using standard ETS procedures for the hand scoring of exercises by non-professional scorers. The scorers were high school graduates.

Those people who worked on the scoring of essay and short answer exercises reported that they were impressed with the unpredictability and creativity of many of the students. Student responses tended to be far more varied and ambiguous than anticipated, and required the redefining and rewriting of scoring specifications to provide for a wider spread of responses which seemed to fit the realm of reasonableness.

Writers of open-end exercises gave only broad general guides to the scoring of such exercises. For example, an exercise (adult level) which requires the person to write a letter of recommendation for an associate gives the following scoring guide:

1 = no response to stimulus
2 = unacceptable response, letter likely to do more harm than good
3 = acceptable response, letter informative and adequately written
4 = as in 3, but with particular appeal and/or conviction

Even though this exercise calls for professional readers, this scoring guide is only the beginning of what is needed (see previous section on scoring). What is the meaning of "likely to do more harm than good," or "informative and adequately written"? The position of the exercise writer has been that responses can only be scored after the assessment has been made and a scale is developed from those responses. During the first assessment National Assessment accepted this point of view but on revision is requiring definition and specific examples of responses at each scale level. Expansion of scoring keys is more time consuming and ex-

pensive, but does provide a better basis for reviewers who are asked to make judgments about the exercises.

The scoring experiences proved to be one of the most valuable aspects of the entire field operation. The implications for the future were clear. Provisions were made for procedures which assured the development of more complete scoring keys much earlier in the developmental process.

Additional selected reactions of the field coordinators follow.

Selected Reactions Offered by Field Coordinators. A vital part of the information which has come to the project has been the "soft data" or the subjective judgments of people who have done the field work.

From the reports made by ETS field workers, the following excerpts have been taken:

1. *On Content*

There was general enthusiastic response on the part of teachers, administrators and directors of teaching to the exercises included in the booklets. Many adults were shocked by the difficulty of some of the items (undoubtedly because they didn't know the answers), but they were also positively affected by the practical problems of reasoning and achievement that did not overly emphasize the recall of specific facts of information.

Quite in contrast to the reaction of some of us at the field coordinators' meeting we are happy to report there was no instance of any educator raising questions about the undesirable or offensive nature of item content. This is a very positive sign and bodes well for the contents that might be used in subsequent activities.

There was no instance of a school administrator or a central office staff person who did not feel that the contents of the tests and the types of items were not important and desirable. Although many felt that some of the materials were extremely difficult for the age group (especially the 10% items which were constructed for success for only 10% of the age group), there seemed to be validation of the appropriateness and relevance of the content to the several areas that were being assessed.

We needed more 'involvement items' in science such as the one asking the kid to write a short paragraph on how he would go about determining whether regular or premium gasoline was needed in a certain automobile. All in all, we thought that there were a great many good items in these exercises. They had much more appeal for the kids than off-the-shelf achievement tests. We are obviously on the right track.

2. *On Tape and Timing*

The male taped voice for the administration of the tests has been positively received in the third and seventh grade classrooms. Timing of the administration is generally seen as good; however, in classes of very advantaged and high-achieving students, the timing tends to be slow, and some students will work ahead beyond the administrator's instruction. Quite in contrast, the timing seems very appropriate for the average and less able student who just has time to mark his responses in the booklet in the time allowed.

In the extremely disadvantaged areas where students were severely retarded in their language and reading development some of the items were extremely difficult, since they were taken by the students in largely a 'listening mode'. Long passages or stories that had to be retained for use with several items which followed were a far more difficult task for such populations than for the students who could make reference back to the material by reading it as subsequent questions were presented.

The taped voice was acknowledged to be good or not objectionable by a great majority of kids. I asked them directly at the end of the session if I had time. Almost all of the kids either enjoyed or accepted having the auditory duplication as they read. Those who could not abide the slow pace worked ahead and finished the booklet. However, some upper kids told me that even though it was too slow it was an interesting change to listen to directions without having to read. The low socio-economic kids really appreciated the taped voice.

The experience of administering the materials was pleasant. I was especially fascinated by the positive way the participants reacted to the tape recorder. Almost all of them indicated that they liked the tape recorder approach rather than have a monotone teacher or counselor read to them. I also noted that comprehension appeared to be enhanced when the children were able to hear as well as see the material presented. This was especially true with children for whom English is a second language.

3. *On the Use of "I Don't Know"*

It was observed that the 'I don't know' category was used in different frequency by students of various levels of ability. In several of the classes of low-ability and disadvantaged-background students, the 'I don't know' category was seldom if ever checked. In contrast, average or higher ability students seemed to be more free to choose this option. It seemed apparent that some students felt it was very undesirable to ever admit that

they didn't know, and thus they would make a guess at one of the four options offered.

4. *On Format and Motivation*

Several of the teachers at the third grade level expressed appreciation for the fact that the booklets contained only one item on a page. They felt this had a very desirable psychological effect on the primary age student who frequently reacts negatively or with despair when confronted with an 8 x 11 page fully packed with 20 or 30 test items. Although this uses a great deal of paper, it was believed this was extremely desirable for the motivational effects on students.

Reactions of Students. In one of the parochial schools in the mid-West, the teacher of one seventh grade class asked her class to write a brief essay on what they thought of the "testing" session.

Selected responses follow:

The test was interesting and I enjoyed it very much but the last part I didn't like, but the rest of it was good, the man spoke very ~~so~~ clearly and his voice I was interesting to hear. The man who came with the test was very nice and said we wouldn't be graded that was very kind of him. This is the first time I have had a test with a person speaking from a tape.

I was scread I don't know what to do. And I thout that everything was wrong or that they might think that I am dum. I don't even know what they word punishment or anything else.

One thing I didn't like about the tapperecorder. It went to slow for me. When the man said stop it scared the daylight out of me. But the man was cute. Nice voice, cool hair, and neat taperecorder.

I think this test we took was very modern and simple unlike the others. The other children asked what the test was for it was almost all scientific stuff. I had some science and I like it but science is nothing but natural common sense. It was fun. Sister Jerome would love to have a tape recorder like that one. It was from Tokyo Japan.

The End

My impression of the test was unusual I thought it was the most easiest test I ever had. The man on the tape said everything in order & correct puntation. But what I didn't like is the way he said "I don't know".

In someways I liked it, and someways I didn't like it. I liked the first part because it was easy. I didn't like it cause of the puns. Still I didn't like it cause the man on the tape was talking to fast and I didn't understand all of the directions. But everything else was ok, I ~~guss~~ guess.

I did not like the test, because the record was fast, and I did not under stand the man ~~when~~ when he was talking about puns and the other stuff, ~~They~~ They want to see how people my age feel about things like that I did not even understand the test.

I liked taking the test especially the puns and jokes. But I want to know how I did on the test

Well I think thats the funnest test I ever took. & One q uestain on every page some of them were easy some were hard. The man was talking to fast and some time to slow

Anecdotes. As with many studies of this nature there were some interesting anecdotes which accompanied the tryouts, not all of which are printable. However, a few examples (in the students' own words) follow.

One enterprising student added the following exercise to his booklet:

I think tests like this are weird because

They are sissy.
I don't get a grade.
Because they represent the establishment.
We never find out the results.
I don't know.

Another youngster offered:

I believe that tests of this nature are biased and are slanted to eliat a response closest to WASP culture. I most sterongly protest the use of tests of this type as they stereotypes mentality of the particepants.

There was the youngster who responded:

Ht: *6-2*, Wt: *178*, Hair: *long*, Eyes: *two*.

Another offered as an example of the kinds of decisions he helps to make:

Well lots of times we can't decide weather we should go pick up girls or go drinking. I help make Hard decisions like these.

Finally, there was the youngster who identified the Speaker of the House as MOTHER.

Individually Administered Exercises

The individually administered exercises were tried out by AIR (ages 9 and 13) and NORC (ages 17 and Adult).

Ages 9 and 13. The procedure followed by AIR involved individual interviews of six students, three boys and three girls, "high," "middle" and "low" ability in each sex. Socio-economic levels were mixed. A Personal Information Form was prepared for each student. Students were selected from four geographical areas, three in California and one in Oregon, and represented a wide range of ability and SES level.

For most questions, half of the students (three) were interviewed by a female examiner and the other half of the students were interviewed by a male examiner. A mixture of boys and girls was seen by each administrator.

Indications of both ability and socio-economic level were relative, within the school or within the district, depending upon both the type of ability measure available and the way in which the general background information was supplied. Records vary greatly from school to school,

and in many cases pupil information was scant or entirely missing. In most cases some background information was gained from a combination of: (1) examination of permanent records, (2) conferences with teachers and principals and (3) conversation with the child himself.

The child was introduced to the task by telling him (truthfully) that he was helping "fix up" a test so it would be good enough to give to other students next year.

The examiner worked with a small microphone taped to the desk or table between the examiner and the child, with the tape recorder more or less out of sight on a different table. This arrangement proved most satisfactory; students were not distracted by the equipment.

The student was told that what he had to say was important evidence to be used in deciding whether any question could be improved or not. The tape recorder was used so the administrator did not have to write all the conversation and so that the tape transcript could be typed and cross-checked against the notes which the administrator did take.

One standard copy was prepared of each exercise along with whatever other visual material the student was intended to see. The administrator gave the student this material, then read the question to him, or played a taped reading, and the student made his answer. Afterwards, the student and administrator went over material which seemed to need clarification and probing questions were asked when necessary.

It was found that the time to administer an item varied widely by student; some gave long answers, some required much probing, others were very quick to respond. The total amount of time spent with a given child varied from 20 minutes to 60 minutes depending upon the specific exercises and the ability of the child to attend to the task.

Specific responses made by the subjects and the observations and judgments of the two field workers were forwarded to the respective contracting agencies which had developed them for their review and consideration. Revisions were made in the light of information gained.

Ages 17 and Adult. Exercises designed for individual administration at age 17 and a selection of exercises written for adults were tried out by NORC in the home using standard interview techniques. The pool of exercises in the five areas designed for individual administration to 17-year olds and adults were assembled into 13 packages of approximately 45 minutes working time each. Two of the 13 packages were for age 17 and the remainder for adults.

The exercises within the booklets were assembled following accepted

survey procedures. That is, each package began with an easy open-end question which almost all respondents could answer and one designed to help the interviewer establish rapport with the respondent. Some exercises were self-administering rather than interview type and these were grouped together and color coded for easy administration. When the interviewer reached the self-administering questions he handed the package to the respondent who read and wrote his own answers to subsequent questions until he came to the end of that section. If it became obvious that a person was having difficulty reading, the interviewer would assist and read the word(s) or question to the respondent. There were a few exercises directed specifically towards employed persons or persons who had children. These were grouped in one of two packages and the interviewer screened specifically to determine whether an adult was eligible to take one of these two packages.

Within the above limitations, exercises were mixed as to subject area and difficulty level. With the exception of a few exercises in the areas of Citizenship and Science, all of the exercises in the pool for administration in the first year of the assessment were tried out. More than half of the literature and social studies exercises were also included in the packages.

The interviews were done in middle class and lower middle class areas (census tracts with average family income of $8,000 per year and $5,000 per year, respectively).

Following the interview a rating form was completed by the interviewer.

For the purposes of the tryouts a 17-year-old was considered eligible for interview if he was born between October 1, 1950, and December 31, 1951. Any adult between 26 and 35 (inclusive) was considered eligible for the adult interview. Several exercises concerned information and opinions about schools. These exercises were all grouped in one package and administered to adults who reported on the screening questions that they had children between six and 17. (This screening was only partly effective; one respondent had children who lived with his ex-wife far from this area so he was unable to answer questions about local schools and school organizations, for example. The screening question should have been "Do you have any children between six and 17 who live here with you?") Another short set of exercises asks about kinds of material the respondent reads or writes as part of his job, so for that package the interviewer screened for employed adults.

As few changes as possible were made in exercise wording, although

the wording of directions was made to make them suitable for the interview situation, instructions were omitted that were given in a preceding similar exercise, and so on. The goal was to test how well the exercise would work in an interview situation with adults and 17-year-olds.

The actual number of individual interviews conducted included 69 adults (33 male, 36 female, 36 white, 33 black) and 13 17-year-olds (seven male, six female, seven white, six black). The interviewers recorded responses to interview-type questions and the respondents recorded their own answers to the self-administered exercises.

The exercise by exercise information together with summaries of the responses and general comments were forwarded to the respective contractors for their use in making revisions.

How Did Tryouts Affect Exercise Development?

The experience with tryouts affected the future of exercise development in a number of important ways.

It confirmed the need to pay closer attention to the development of more adequate scoring rationales and very detailed and specific scoring keys. The problem of scoring had been noted earlier in other specific research studies and in problems encountered in review conferences. Tryout confirmed these earlier experiences and pointed out a major flaw in the developmental process.

These massive tryouts pointed the way towards using less extensive tryouts in the future. A large scale tryout was beneficial to gain information on general problems which might occur in administration under conditions which simulated the actual assessment, and information helpful to the refinement of exercises was obtained. However, much of the same information could have been obtained by careful sampling of exercises to be tried out. In many subject areas there are "families" of exercises which are similar. Information gained by trying out one or two exercises in a "family" could be generalized, in many cases, to other exercises in that "family."

Finally, less emphasis on measuring difficulty levels and more emphasis on obtaining feasibility type data has been another direct result of tryouts. Some indication *is* needed that there are enough exercises in the three categories (easy, average, difficult), but beyond that, precise information is not necessary as it is obtained in the actual assessment. Obtaining feasibility data which is designed to improve the understandability and reduce the vocabulary level of the exercises *is* important in the preassessment tryouts.

CHAPTER VII

Final Reviews and Selection

In addition to the reviews which have been described in the preceding chapters, three additional reviews preceded the selection process and helped determine the actual exercises to be used in the first assessment. They included:

1. A review by lay persons to judge the *meaningfulness* of the material,
2. A review by the Technical Advisory Committee (TAC) for ambiguities and clarity of wording,
3. A review by the United States Office of Education (USOE) for potential invasion of privacy.

The same five areas which had been through tryouts (Citizenship, Literature, Science, Social Studies, Writing) were taken to the lay group, while, due to time pressures, only those areas to be included in the first assessment (Citizenship, Science, Writing) went to TAC and USOE.

The actual selection of exercises to be used in the first assessment was made at a conference of special consultants, contractors and staff and will be described later in this chapter.

Review of Exercises for Meaningfulness

As work continued on the refinement of the exercises, the staff realized that there were some exercises which asked rather unimportant information or were trivial. For instance, the following exercise was written to assess the understanding of 9-year-olds of some of the major relationships involving culture, the group and the self and, in this instance, the importance of the family:

What are some things almost all American families do together?

This particular exercise had been worked over extensively following comments from the mail reviewer and from both subject review conferences. The reviewers believed it was valuable to assess the role of the family in American culture but that there must be a better way of doing

it. In its present form the exercise is confusing. Does it ask what *individuals within the family* or *one family and other families* do together? What would the responses show? What information would be obtained from such an exercise?

The recognition that some trivial exercises existed led to still another review of exercises. Again the lay people were called on to provide some direction. The conferees were from 17 states and represented many of the same organizations that were represented at earlier lay conferences. At a conference held in Chicago in March 1968, 18 lay people met to consider the meaningfulness or importance of each exercise in a 20 percent sample of assessment exercises in Citizenship, Literature, Science, Social Studies and Writing. The staff had selected approximately 10 percent of the exercises as ones which in their judgment were the least meaningful. The other 10 percent were selected at random. The conferees were asked to judge each exercise on whether it was meaningful and important to ask. These five subject areas were chosen because they were the ones in which the developmental work was completed or nearly completed. The other five areas, at this stage, still needed considerable work.

Purpose

As each exercise in the assessment is to be reported individually and not as part of a battery of exercises, each one must have content validity. Evaluation of content validity varies, of course, depending upon the knowledges and skills of the person looking at the exercise. A chemist can look at an exercise involving a chemical formula and judge whether it "makes sense," whereas a person who has never studied chemistry would have no basis for a valid judgment of whether or not it makes sense. The reactions of subject matter specialists and other educators to the exercises had been obtained at subject matter review conferences. This conference was organized specifically to obtain the reactions of lay people to the meaningfulness and importance of the exercises.

Organization of Conference

The 18 conferees were divided into five groups. On the first day, each group met with a member from the staff and considered about 100 exercises. Each group had different exercises. If the group had a question about an exercise, it was put aside to be presented the next day to the entire conference. If only minor editorial changes were suggested, these were noted and were later referred to the appropriate contractor for consideration.

Results

Of the 468 exercises taken to the conference, there were 93 which had been questioned by one of the groups, and hence, were presented on the second day to the assembled conferees. Discussion followed and an attempt was made to judge the consensus of the group. Of the 93 exercises, only 36 were considered trivial enough to warrant a recommendation to shelve.

Table 7-1 shows the number of exercises selected by the staff as least meaningful and the number chosen at random for each of the five subjects.

TABLE 7-1
Lay Conference on Meaningfulness
Number of Exercises Chosen by Staff
March 1968

	Criteria for Selection		
Subject	Staff Selection as Least Meaningful	Randomly Selected	Total
Citizenship	32	18	50
Literature	43	63	106
Science	88	64	152
Social Studies	51	73	124
Writing	22	14	36
Totals	236	232	468

No attempt was made to select equal or proportional numbers of least meaningful exercises from each of the five subject areas since there were wide differences in both the total number of exercises in the respective subject pools and in general content. Rather, the approach was to select across all areas those exercises which were least meaningful in the judgment of the staff. This total was to represent about 10 percent of the number of exercises in the combined pools. The number of exercises chosen at random then supplemented those purposefully selected to total approximately 20 percent within each subject area.

Table 7-2 shows by category the number of exercises questioned by the original small groups and referred to the entire conference the next day for further review.

Of the total 236 exercises selected because of questionable meaningfulness, 55 or 23 percent were also questioned by the small lay groups. Of the 232 randomly selected exercises, 38 or 16 percent were questioned by the small lay groups. These are overall totals for the five subject areas combined. If the figures are compared within each subject area it will be noted that only in Writing was there a definite tendency for the lay groups to question exercises purposely selected by staff as compared to those randomly selected (41 percent vs. 0 percent). In all other areas there were slight tendencies for the lay groups to question purposefully selected exercises over randomly selected exercises, but the differences

TABLE 7-2
Lay Conference on Meaningfulness
Number of Exercises Referred by Individual Panels to Entire Conference
March 1968

Subject	*Criteria for Selection*					
	Least Meaningful			*Random*		
	Original N	Questioned by panel	%	Original N	Questioned by panel	%
Citizenship	32	7	22	18	3	17
Literature	43	9	21	63	11	17
Science	88	15	17	64	9	14
Social Studies	51	15	29	73	15	21
Writing	22	9	41	14	—	0
Totals	236	55	23%	232	38	16%

are not great (17 percent to 14 percent for Science, 22 percent to 17 percent for Citizenship, 21 percent to 17 percent for Literature, 29 percent to 21 percent for Social Studies). As the lay reviewers had questioned the meaningfulness of a substantial number of the exercises submitted to them, TAC included in their final review (see next section) an emphasis on the meaningfulness of each exercise.

Table 7-3 shows the recommendations of the conference: the number of exercises to be shelved, the number needing additional editorial work and the number approved.

TABLE 7-3
Results of Lay Panel Conference
March 1968

Subject	No. Exercises Approved	No. Exercises to be Edited	No. Exercises Recommended to be Shelved	Total No. Exercises
Citizenship	40	10	0	50
Literature	87	7	12	106
Science	134	10	8	152
Social Studies	95	18	11	124
Writing	25	6	5	36
Totals	381	51	36	468

An example of the type of exercise which was shelved at the conference on meaningfulness is the following one written for 9-year-olds to see whether they knew some of the fundamental facts and principles of science:

Directions: Choose the best answer.

Suppose you have several crystals. When you look down on the tops of them, you see the following:

I II III IV

Which two are most likely to be the same substance?

(A) I and II
(B) I and III
(C) II and III
(D) III and IV
(E) I don't know.

The lay reviewers felt this exercise could be easily guessed and hence would have little meaning as a measure of knowledge of principles of science.

Another exercise which was considered unimportant, or trivial, was the following one designed to assess whether the 9-year-old had an accurate perception about scientists:

Do you think that all scientists wear uniforms?

(A) Yes
(B) No
(C) I don't know.

The panelists responded, "Who cares? It is unimportant."

As may readily be seen, the potential loss exemplified by the recommendation to shelve 36 exercises was minimal. Within subject areas there is an indicated loss which ranges from "no loss" (Citizenship) to 14 percent loss (Writing). Reasons for shelving exercises included "much too obvious," "would get the right answer for the wrong reason," "so what?" "a simple aptitude item" and "unimportant to know."

Conclusions

By and large the conferees felt the exercises were meaningful and were important to ask. They were least critical in the areas of Science and Citizenship. An admitted lack of knowledge about science and a willingness to accept the subject matter specialists' judgment on the importance of these exercises contributed heavily to the acceptance of Science. No recommendations were made to shelve any of the citizenship exercises, which may reflect the extensive work of earlier lay panels with citizenship exercises in the developmental stage.

The group was less willing to accept the importance of some of the literature exercises, perhaps because they felt better acquaintance with the subject.

Final Review by the Technical Advisory Committee

The Technical Advisory Committee (TAC) for National Assessment, composed of a group of men eminent in the fields of psychological measurement and statistics, had been very active and influential in guiding the technical aspects of the project. As the time for the assessment drew near, members of TAC expressed their concern and feeling of responsibility by performing one final review of all materials prepared for Citizenship, Science and Writing before the final selection of exercises to be used in the assessment was made. This review was meant to supplement those made earlier by subject matter specialists, educators and lay persons. TAC continued to stress sensibleness, meaningfulness and clarity of wording as essentials. They were particularly sensitive to ambiguities and legitimate misinterpretations and strove to first detect and then change or delete them. Some of these ambiguities and legitimate mis-

interpretations could actually place bright, capable youngsters at a disadvantage.[34]

The committee met in July 1968. All exercises in the citizenship, science, and writing pools as of that date were read independently by two TAC members. Any exercise that received a questionable response from one reader was reviewed by the total group. Exercises were then accepted, revised or shelved.

A side benefit from this review included some rather specific recommendations which emerged with regard to guidelines for the future and for general format and style. Although some of the points may seem trivial to the reader, they represent an emerging style for National Assessment. As such they are of historical interest to all and may be of specific assistance to others who have asked for details regarding the National Assessment model.

From the first review the following recommendations were made:

1. The term "one" should be added to all multiple-choice exercises that read, "which (one) of the following . . . . "
2. Exercises that are designed with negative stems (i.e., "which one is not" or "all of the following EXCEPT") should be revised if at all possible to eliminate the negative or exceptive approach.
3. All exercises should be reviewed by a grammarian.
4. Long written exercises should be physically arranged on pages so that the question and the space for writing are adjacent to each other, with the stimulus on the left-hand page.
5. Care should be taken to avoid over-using any given response position. A random arrangement of choices was recommended.[35]
6. For exercises using science apparatus an attempt should be made to judge quality of performance as well as specific knowledge or action.

[34]Banesh Hoffmann published *The Tyranny of Testing* in 1962, in which he condemned the usual multiple-choice type test items as favoring "the superficially brilliant and penaliz[ing] the student who has depth, subtlety, and critical acumen." He has been particularly critical of items which are open to misinterpretation by bright students, who, because they have more information, are likely to give an answer considered incorrect by the writer, while those with less ability and information are able to give the expected answer.

[35]Staff subsequently devised a system which would randomly distribute the correct answers. For one- or two-word responses the choices should be arranged in alphabetical sequence. For longer choices arrangement by length (shortest to longest) was recommended unless there is some other logical sequence.

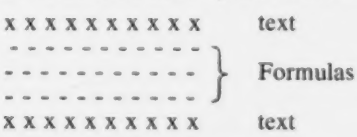
7. Exercises which TAC shelved but which seem to be quite innovative should be looked at again to see if they can be salvaged. Perhaps they should be tried in order to avoid establishing a National Assessment stereotype of staid, traditional questions.

In a subsequent review of mathematics and music exercises held in November 1968, further general recommendations emerged including:

1. The use of "I don't know" as an option should be used routinely with all multiple-choice exercises with the exception of rare instances.
2. The use of "none of these" as a choice should be extremely limited and if it is used it should occasionally be the correct answer.
3. Where choices consist of a set of diagrams or pictures, the response boxes should accompany each diagram rather than be listed separately.
4. The choices of multiple-choice exercises should not be numbered or lettered, just listed together with a place for the student to mark.
5. All graphs should carry either a self-explanatory title or some introductory statement.
6. Exercises which deal with current issues may cause special problems due to lapse of time between their development and their use and should be used sparingly and only if they can be updated at the time of assessment. Exercises of a political nature are potential problems. Exercises dealing with postage rates and money exchange are also susceptible to problems of change over time.

Recommendations specific to the area of Mathematics included:

1. Be alert to the proper usage of "equivalent" vs. "equal." "Can be written as" can be used in place of "equals."
2. "Best estimate of" should be replaced by "is closest to."
3. Where a number of formulas are given in the stem, the formulas should be set in individual lines by themselves. That is:



4. Choices which list number of inches, feet, minutes, etc., should be so labeled. That is:

30 minutes
45 minutes

60 minutes
90 minutes

I don't know.

5. Mathematics exercises are inclined to be sex-biased in favor of males. An effort should be made to frame exercises in terms familiar and interesting to females.

6. Some consideration needs to be given to the presentation of mathematical symbols if the exercises are to be taped in the normal routine fashion. For example, how is a question presented whose alternates are:

(As of the writing of this monograph, this last issue had not been settled.)

Having established these general guidelines, TAC expressed the wish that the final review function be assumed—for later subject areas—by a special TAC sub-committee. Their suggestion was implemented and a group was formed which later became the nucleus for the Exercise Development Advisory Group (EDAG). Represented in this group are specialists in educational measurement, a political scientist and educators, two of whom have extensive experience with low achieving students and inner-city problems.

EDAG subsequently continued to act as a final review board for exercises in other subject areas. In addition, their function has broadened to include the consideration of general policy problems relating to the development of exercises and of related special problems.

USOE Review for Invasion of Privacy

All of the research and development which had gone into the preparatory phase of National Assessment (1964 to 1968) had been funded through private sources. As the time neared for the first actual assessment, additional funds were needed, and the foundations which had been supporting the project expressed their wishes that as soon as possible the operational phase should be supported by public funds. Federal funds did subsequently become available and the USOE began taking an active interest in the project and providing an appreciable part of the support.

One of the results of this shift in financial support came in August 1968, when all exercises in the citizenship, science and writing pools were reviewed by the Chairman of the Internal Clearance Committee of the Bureau of Research, USOE.

The principal purpose of this review was to judge whether any of the questions might be interpreted as an invasion of privacy, this issue having been a particularly sensitive one in recent years.

As a result of this review involving approximately 1,000 exercises, four exercises were removed (one science and three citizenship) and 11 were modified (all citizenship).

The exercises which were removed, together with selected comments, follow:

Age 26-35, Science:

> How are the components of contraceptive pills which contain estrogen and progestine intended to function to prevent conception?
>
> By killing the sperm cells
>
> By preventing the ovary from releasing an egg cell
>
> By speeding the passage of unfertilized egg cells out of the body
>
> By preventing unfertilized egg cells from passing out of the body
>
> By causing the ovary to release the egg cells prematurely, before they are capable of being fertilized
>
> I don't know.

There are two phrases in the stem of this exercise which [offend some religious denominations]: namely, "contraceptive pills," and "prevent conception." . . . Less objection can be taken to the phrase, "contraceptive pills"—a phrase which is, after all, in reasonably common usage. —Our view is that it might be better to omit this item, unless it really makes a uniquely and indispensably significant contribution to measurement of the adults' knowledge of science (which does not seem very likely). Especially in the first year of the Assessment, when critical attention to the Assessment is likely to be most widespread, cautionary discretion would seem to us very desirable.

Age 26-35, Citizenship:

> Have you ever been called to school to talk over any misbehavior of your child?

This exercise is open to two objections: (1) it does not, on its face,

seem *directly and clearly related* to the citizenship of the respondent (i.e., the parent); and (2) it may be easily viewed by the respondent as excessively "personal," an unjustifiable "prying" into family affairs. Also—from our viewpoint—it is unfortunately highly quotable by all those (a goodly number, in recent years) who assert that the questionnaires used by researchers too often represent "fishing expeditions," with little regard to the feelings and privacy of the respondents. We would suppose that this exercise can be eliminated without excessive damage to the measurement of parent's citizenship.

Age 26-35, Citizenship:

> During the past year, have you been attacked physically by anyone?

Like the preceding exercise, this one does not, on its face, seem directly and clearly related to the citizenship of the respondent. Moreover, in the case of women, the phrase, "attacked physically," is likely to be interpreted as a polite expression for "attacked sexually," or "raped"—and certainly this is not the kind of question that (from the viewpoint of respect for privacy) should be asked by any but a privileged person, and even then only in special circumstances.—Our judgment is that the two objections—namely, lack of direct and clear relevance to citizenship, and the likelihood of offending female respondents—make this particular exercise unusable in a national survey.

Age 9, Citizenship:

> Do you have a T.V. at home? (If yes) Do any of these men [in a set of 8 photos] report daily news on television? (If yes) Point to one of them.

The first of the 3 questions in this exercise may be shame-generating in the case of a 9-year-old, when the absence of a TV-set is due to poverty. This difficulty can be readily obviated, however, by substituting the question: *"Do you watch TV"?* This revised question seems to us more pertinent than—or at least as pertinent as—the original question; and the substitution of this question for the original is hereby recommended.—If the idea of TV *in the home* is considered essential, then a more diplomatic or considerate inclusion of this concept may be effected by the use of two questions rather than one; thus:

> Do you watch TV? (If yes) Do you generally watch it at home?

The Selection Process

After the final reviews, there was the matter of determining exactly which exercises were to be used in the first assessment. Financial considerations did not allow unlimited use in the first assessment of all exercises which remained in the pool after the last reviews. In addition, there was wide variation from one subject area to another in the amount of material which had received final approval by TAC. For example, at age 13, the following number of exercises and their time requirements remained:

	No. Exercises	*No. Minutes*
Citizenship	112	181
Science	195	348
Writing	48	224

It was necessary, therefore, to decide just what coverage was feasible, taking into consideration the funds available. Based on estimates of costs for administration the decision was made that a total of 480 minutes per age could be assessed for all three subject areas. This was an average of 160 minutes per subject area per age. The actual selection of exercises to be used in the first assessment was made at a conference composed of staff and contractor representatives.

Prior to the selection conference, each set of exercises was read and rated independently by two special staff consultants and by a contractor representative for both agencies involved in the development of the material (American Institutes for Research and Educational Testing Service). Each rater was asked to consider these criteria in making their selections:

1. Give first consideration to the quality and reportability of the exercises.
2. Select exercises totally approximately 160 minutes of administration time per subject area for each age.
3. Select a balance of exercises so that each objective is represented and there is an adequate number of easy, average and difficult exercises, but select proportionately more easy exercises.
4. Give some preference to overlapping exercises (that is, exercises which occur at more than one age level).

The indepedent judgments obtained were pooled and those exercises which were ranked high most consistently were considered to be "pre-

liminary selections" and were grouped together to form a nucleus from which the conference participants could work.

The first Exercise Selection Conference, for ages 17 and adult in Citizenship, Science and Writing, was held in October 1968. Three groups consisting of staff, special consultants and contractor representatives met independently (one for each subject area) and made the final judgments as to exactly which exercises should be used in the first assessment. A similar procedure was followed in February 1969, to select exercises for ages 9 and 13. When preference was given to exercises which overlapped with those already selected for ages 17 and adult, a major part of the selection for ages 9 and 13 was already done. For this reason, the second selection conference was held in one day and performed by one staff member working with three special consultants.

This method of selecting exercises proved satisfactory and plans were made to follow a similar scheme in subsequent years with one modification. That modification was the addition of special subject matter consultants, so that for the second assessment year the group selecting Literature (ages 9 and 13) included a specialist in that field and the group selecting Reading (all ages) included a specialist in reading.

Summary

The final steps of preparation for the first assessment included three special reviews which preceded the process of selecting actual exercises to be used in the assessment.

1. A review by lay persons was made in an effort to evaluate the meaningfulness of the material and to exclude materials which seemed trivial in nature.
2. The Technical Advisory Committee reviewed the materials in an effort to eliminate ambiguities in the exercises.
3. The United States Office of Education reviewed the materials and suggested eliminating or changing exercises which might be construed as an invasion of privacy.

A final selection of specific exercises to be included in the first assessment was made at a conference which included staff, special consultants and contractor representatives.

CHAPTER VIII

New Directions in Exercise Development

What has come of the experience gained in the early phases of the project? What has been learned and what are the future directions of exercise development at this stage of experience? On the basis of the previous five years of experience a more structured plan for developing National Assessment exercises has emerged.

One of the considerations at this time is whether the project should continue to have materials developed by contractors or whether materials should be developed under the direct supervision of National Assessment staff. In the beginning it was an absolute necessity that the developmental work be contracted out. The staff was small and the future uncertain. A number of test construction agencies existed with highly competent and experienced staff, and it made sense for National Assessment to use and benefit from the expertise which already existed.

However, a number of problems in the early developmental process have been identified. In a number of instances, the contracting agency assigned only one writer to the job of writing all materials. In some cases this tended to produce materials which were lacking in variety and interest compared with those areas where a number of writers were involved.

Another problem was that there did not seem to be enough checkpoints along the way. The contractor was awarded the contract and exercises were not reviewed until all were written. This gave staff and reviewers no opportunity to evaluate and redirect the efforts of the writers, if this was necessary. As a result, some subject areas suffered heavy losses through the review process, and it became necessary to prepare additional materials to add to the remaining pool of exercises. When this happened, the staff invited independent subject matter consultants to develop new exercises for sections considered by the reviewers to be inadequately covered. Actual exercise writing conferences were held in some cases, and groups of specialists working with staff spent two or three days developing new materials. The results of using in-

dependent subject experts were so rewarding that staff began to consider the possibility of doing the developmental work for one or more areas.

The first step in this direction was taken in the summer of 1969, when it was necessary to begin the revisions of Citizenship, Science and Writing. The decision was made to have the contractors continue working in the areas of Citizenship and Writing and for the National Assessment staff to undertake the revision of Science. The second step was taken in November 1969, when the decision was made for staff to direct the revision of Reading for the second cycle. In both Science and Reading, special subject matter consultants have been enlisted to guide and prepare the revisions of objectives and the new exercises, while staff continues to function in a coordinating role.

The extent to which this trend will continue is unknown at this time. It seems reasonable to expect with greater staff involvement there will be greater opportunity to try out ideas. Also, problems of communication which seem to be inherent when working with large organizations at widely separated distances should be lessened. A better evaluation can be made after more experience is gained.

Possible disadvantages of developing materials under direct supervision of staff relate mainly to the danger of "in-breeding" or restricting the viewpoint to staff and those persons who actually prepare the exercises, thus destroying the monitoring role of staff. In order to lessen this possibility, a plan has been designed for the development or revision of areas which not only incorporates all the previously used review and field testing procedures, but also adds additional safety features designed to assure broad representation of varying educational, subject matter and lay viewpoints. The general plan which follows was based on five years of experience. It is used by contractors as well as by staff in the supervision of staff developed areas.

Five Phases

The developmental process has been divided into five logical phases:

1. Phase A. Development (or revision) and review of objectives and prototype exercises.
2. Phase B. Preparation of exercises.
3. Phase C. Review and revision of exercises.
4. Phase D. Field testing and revision of exercises.
5. Phase E. Final reviews and selection.

The amount of time devoted to each phase varies with the subject area. This is necessary because of the current cycling plan which calls for three areas (Reading, Mathematics, Science) to be assessed every three years. Other areas are to be assessed every six years. Specifically, the schedule is as follows:

Year		*Subject Areas*
01	1969-70	Citizenship, Science, Writing
02	1970-71	Literature, Reading
03	1971-72	Music, Social Studies
04	1972-73	Career and Occupational Development, Math, Science
05	1973-74	Reading, Writing
06	1974-75	Citizenship, Art

Some deviation from the three-year, six-year pattern was necessary in order to establish the first sequence. From year 07 on, it is now planned that all areas will be spaced at either three-or six-year intervals.

From the developmental point of view, six years allows a far better schedule to be prepared in that it permits additional steps to be built in if that becomes necessary. Past experience indicates that schedules are difficult to keep if unexpected problems are encountered. Three years is a minimum amount of time needed to develop an area and allows no safety margin of time if unusual problems occur.

Subject areas on three-year cycles have an additional disadvantage in that their revision must begin before results from the previous assessment are available. This is unfortunate, but results can be assimilated as they become available during Phase B.

For those areas on three-year cycles the time allotment is as follows:

Phase	*Number of Months*
A. Development (or revision) and review of objectives and prototype exercises.	9
B. Preparation of exercises.	9
C. Review and revision of exercises.	6
D. Field testing and revision of exercises.	6
E. Final reviews and selection.	6
Total	36

Time allotments for areas on six-year cycles are increased proportionately, and revisions are not started until results from the previous assessment are available.

For areas on a six-year cycle the time allotment is as follows:

Phase		*Number of Months*
A.	Development (or revision) and review of objectives and prototype exercises.	14
B.	Preparation of exercises.	14
C.	Review and revision of exercises.	10
D.	Field testing and revision of exercises.	10
E.	Final reviews and selection.	10
	Total	58

While even the three-year period may seem long, considerable detail is required to prepare materials, plan and conduct reviews and tryouts and then prepare revisions. The total process will be explained in the sections which follow.

Objectives and Prototype Exercises

Phase A allows for the reconsideration and possible revision of objectives in an old area, if this is necessary, or the development of objectives in a new area. While all three of the year 01 subject area objectives are currently going through revision, it is possible that not all areas will need revisions. At least the first steps described below will be taken for each old area. If, however, it appears that little or no change has taken place in the curricular emphasis since the last revision, no changes in content would be made in the objectives. It is likely, however, that some editorial changes may take place in an attempt to arrive at a more uniform format for objectives in all subject areas.

A variety of input is sought in the early stages of developing or revising objectives. A review of the literature is made, in order to determine what current thinking is among curriculum specialists. In the case of a revision, the question becomes, "What of consequence has happened since the last revision?" and "To what extent do the existing objectives reflect current thought in the field?" In addition to a review of the literature, other sources may prove helpful. Many state departments of education and school districts have excellent materials. Recently an Objectives Exchange has been established at the University of California at Los Angeles. The Exchange is building a library of educational objectives

stated in behavioral terms which have been collected from schools throughout the United States. As their library is enlarged, this source is expected to provide valuable information.

In the case of a revision, a number of specialists are consulted and asked to critique the existing objectives. People asked generally include some who have been previous reviewers for National Assessment and some who are new to the project. A broad representation is sought. An attempt is made to select people from elementary and secondary schools (public, private and parochial); people from colleges and universities; persons knowledgable about low socio-economic problems and inner-city problems; members of minority groups (Black, Mexican-American and others); people from different geographic areas (Northeast, Southeast, Central, West) and students. A total of 10 to 15 people may be asked to do this type of review.

From these various inputs, a revised set of objectives is drafted. If the inputs indicate relatively little or no change has taken place, the objectives may remain unchanged in content. If the inputs reflect major changes in curricular emphasis, or if major omissions are discovered in the original objectives, major revisions will be made.

Unless the objectives remain unchanged, the next step is to hold a review conference for the purpose of critiquing the reformulated objectives. Again, broad representation is sought in the same general categories of people who prepared independent critiques. The main advantage of holding such a conference is that it allows for interaction among the people participating. If differences occur (and they always do), those differences generally can be resolved. Based upon the advice received, the objectives may be further revised or refined.

In the process as described above, both subject matter specialists and educators have been consulted. The next step is to seek the reactions of lay groups. Generally, when a conference of lay persons is arranged, more than one subject area is ready for review. In the summer of 1969, for example, the objectives for the three areas in the first assessment (Citizenship, Science, Writing) had all been revised and were ready for review by lay groups. A conference was held with 35 lay persons, including five students. This was the first attempt to involve students, and they made interesting, stimulating and worthwhile contributions. (The students were in the 17-18 age range; four of them were college freshmen and one was a high school senior.) On the first day of the conference, the panelists worked in seven independent subgroups, each one chaired by a lay person and with a National Assessment staff member present

to serve as a recorder. Each group reviewed the objectives for all three areas. Contractor representatives for Citizenship and Writing as well as the National Assessment science consultants were on call to answer any questions which arose. On the second day the lay chairmen met as a group, together with contractor representatives, the science consultants and staff, in order to communicate their ideas and resolve any differences.

Following the lay review, complete and detailed minutes from both days were compiled and used as a basis for further refinement of the objectives.

The pattern used in the summer 1969 conference just described has proved satisfactory and probably will continue.

If revision is called for after the conference, final acceptability of the objectives is determined by asking a group of subject matter specialists and lay persons to review them by mail. Both persons new to the project and those who have been previously involved are asked to make this evaluation.

Prototype Exercises. While the objectives are being developed or revised, prototype exercises are also being developed. The preparation of prototype exercises is one of the most important preliminary steps in the development of materials. If the area is contracted out for development, the prototype exercise helps assure communication between the staff and the contractor as to what type of material is to be produced. If independent consultants are used to produce exercises under direct staff supervision, the same kind of communication is necessary. In the past, large segments of materials have been produced and shelved. Losses occurred because there were not enough checkpoints in the process. There should have been more opportunities for staff and reviewers to interact with exercise writers before all materials were produced. The process of developing and reviewing prototype exercises assures better communication and less wasted effort.

A prototype exercise is one which is to serve as a model for the development of other exercises in the total pool of exercises. As such, it must be a concrete example of an exercise and as complete in every detail as possible. It must be completely classified as to objective, age, difficulty level, type of administration, time estimate, etc. The exercise should be clearly stated and followed by directions to the administrator if necessary, a rationale for the scorer, directions to the scorer, specific acceptable and unacceptable responses and a scheme for reporting the results. While prototype exercises are not new to the developmental process, earlier examples were not so detailed.

In most cases it is necessary to do some preliminary field testing to obtain specific examples of acceptable and unacceptable responses for open-end exercises or to obtain samples if the responses are to be assigned scaled values (as in Writing).

In order to get ideas for the preparation of prototype exercises, whether contractor prepared or consultant prepared, a conference is held. Subject matter specialists representing different specialties within the field and people experienced at different levels of performance are invited to a conference.

For example, when the science prototype conference was held, 17 people were invited who were specialists in chemistry, physics, astronomy, earth science and biological science. Among them were elementary and high school teachers, a junior high school curriculum specialist, a person who has done much work with the manipulative aspects of science and one specializing in the social implications of science.

The participants spent two days working singly or in small groups writing exercises. The ideas generated by this group were then refined by the special science consultant and prepared for the next step in the process.

Using the group approach to the development of prototypes provides two important things. First, it tends to generate more ideas and a broader coverage of the area than would be the case if only one or two people did the job. Second, it creates a pool of individuals who may be called upon to develop the actual exercises once the prototypes are approved.

To help prepare people to participate in such a conference, a set of "Guidelines for the Development of National Assessment Exercises" was developed. These "Guidelines" discuss specifications, general format considerations and the development of scoring rationales, and give examples of inappropriate exercises. Excerpts from the "Guidelines" follow:

General Considerations

The exercises developed for National Assessment differ from those developed for standardized tests. The Assessment is interested in levels of performance of groups of people in various subject areas. It is not concerned wtih an individual's accomplishments except as part of the total group results. The assessment is interested in reporting such things as "90 percent of all 17-year-olds know the name of the Governor of their state but only 20 percent know the name of their

Representative in Congress." The results of the assessment will be reported in terms of individual exercises.

Since each exercise will be reported, it is extremely important that each exercise has content validity. It must attempt to measure the objective for which it was written. It must also measure something which will be meaningful to report and which will not be considered trivial in nature by the profession or the general public. The assessment results will be reported widely in newspapers, magazines and professional journals and on TV and radio. It is important that each exercise be in no way offensive or liable to a charge of invasion of privacy.

The objectives in each subject area were carefully developed to assure that each objective met the following criteria: considered important by subject matter specialists, accepted as an educational task by the school, and considered desirable by thoughtful lay citizens. In developing exercises effort should be made to see that each objective is sampled.

Criteria for Exercise Development

1. *Creative and interesting.* Be imaginative! Often situational exercises (those which have a "real life" setting) have more meaning and appeal to the person trying to respond to the question. The use of tapes, film strips, discussions, interviews, or actual experiments may be more productive than the usual paper and pencil exercises. Do not feel bound by what is considered "traditional" test item types (e. g., isn't there a more interesting way to present material than the traditional overworked multiple-choice exercise?). Try to produce the best plan for developing an exercise to meet the objective.

2. *Wide range of stimulus materials.* Try to break out of the traditional middle-class bias in educational materials and provide some exercises dealing with Black culture, Black history, Black literature. This could be done by either using stimulus material or referring to authors, composers, artists, etc. Other exercises might relate to the problems of the ghettos, of student rioters, of Indians, or immigrants or of other cultural groups.

   Be alert to possible sex bias. Occasionally exercises seem slanted

to masculine interest or knowledge. If this occurs try to think of other exercises which are equally appropriate for girls.

In some subject areas, materials from everyday experiences are particularly effective. For instance, in Science small experiments with household materials such as water, salt, soda, vinegar, etc., might prove interesting to the younger age groups. In Reading, labels from household items such as canned goods, box tops, newspapers, etc., provide good stimulus material.

Asking 9-, 13- or 17-year-olds for suggestions of things which interest them or which they think are important might be a source for appropriate exercise ideas. Talk to the kids and try to understand what their world is like.

3. *Measurement of higher cognitive levels.* Try to measure the higher cognitive levels, such as reasoning and thinking logically, drawing inferences, reaching conclusions, analyzing and synthesizing different points of view.

   In the area of Reading, exercises of this type might deal with two or three different newspaper articles on one topic, or with a newspaper story vs. an editorial on the same subject. In both cases, comparisons could be asked for and interpretations made.

4. *Either group or individual administration.* As assessment exercises do not have to be administered in a classroom setting, greater freedom in type of administration can be used. Exercises may be group or individually administered. Most people are familiar with the usual group administered test which uses the typical objective questions. But also consider types of situational tasks which might be used with groups. For example, one 45-minute exercise could be given to a group of eight 17-year-olds (or 13-year-olds or 9-year olds). The purpose of such an exercise may be to measure the extent to which the youth constructively contributes to group efforts. Achievement of these objectives is to be inferred from observation of specifically defined behaviors of all members of the group. Behaviors to be observed do not include all relevant behavior exhibited in this type of group situation, but only those behavior categories which extensive tryouts have indicated can be reliably observed.

5. *Wide range of difficulty for each age.* As the assessment is interested in the educational accomplishment of a group of people, it must be able to sample what most of the group can do as well as what very few can do. Thus exercises must be developed which are easy, average and difficult. For convenience we designate these as 90 percent, 50 percent and 10 percent exercises, although in essence a 90 percent exercise may be one which 80 percent to 95 percent of the students can pass, a 50 percent exercise one which 40 percent to 60 percent can pass, etc. (If there are a number of possible responses to one exercise, or some other reason why it would be difficult to indicate the difficulty level, a "multiple difficulty" designation can be indicated. . . .)

   Experience with initial tryout studies of existing assessment exercises has shown that it is extremely difficult to produce an easy or "90 percent" exercise. [One reason for this is that writers tend to be too wordy and fail to use simple vocabulary.] Frequently the vocabulary is too difficult for the student from a lower economic area. Recent research has indicated that difficulty level can be manipulated by the choice of distractors. Care should be taken not to create "easy" exercises by using ridiculous distractors.

6. *Simple vocabulary and clearly stated exercises.* The exercise must be written in such a way that the individual will understand what he is expected to do. The assessment is concerned with whether the individual can do the actual task asked of him, not whether he can do mental gymnastics. This means that the directions and format which go with the exercise must be simple enough for the least able student to understand what is wanted, and the vocabulary, phraseology and length of sentences must not be confusing.

7. *Directionality.* Each exercise must have a "correct" or "desired" response. The purpose of National Assessment is to "take a reading" of desired knowledges, skills and attitudes, and then to reassess the same attributes several years later in order to measure progress over time. Hence every exercise must have a "desired" response. Otherwise, changes that take place over time are neither correct nor incorrect, desirable nor undesirable and tell nothing about the extent to which an objective of education is being attained. Survey-type questions which do not have a correct answer are not acceptable.

8. *Potential problems*

Scoring and reporting. A potential problem in attempting to be imaginative and "break loose" from typical multiple-choice format is that exercises may be difficult or impossible to score and report. Samples of exercises which have been found inappropriate for these reasons are attached.

Timeliness. Exercises which are timely may have considerable appeal. However, give consideration to the effect of the time lag between writing the exercise and when it is scheduled to be used. An example of this is a social studies exercise which requires the student to identify a picture of the Vice-President of the United States. When the exercise was written the Vice-President was Humphrey. This exercise can be easily changed if and when it is used. Other types may not be timely three to five years hence.

Avoid clustering. Writing clusters of exercises based on one stimulus is, in general, to be discouraged. However, exceptions to this may occasionally occur in areas such as Reading or Literature where fairly long passages may be used.

Number of exercises. The time restrictions under which we are currently working allow approximately 240 minutes per age per subject area in the actual assessment. While it is desirable to have a pool in excess of this (500 minutes as a rule of thumb), it is not necessary to write unlimited numbers of exercises. As exercises are produced there needs to be some consideration of time requirements and coverage of objectives.

9. *Coding of exercises.* For each exercise, indicate the objective(s) being measured, the estimated difficulty level (10-50-90-M),[36] whether it is to be group or individually administered, the age level for which it is intended and the amount of time which should be allowed for students to respond. If it can be used at other ages, with or without different difficulty levels, indicate that also. If an exercise needs an individual interview to clarify the task or to probe for certain responses, the exercise should be designated for individual administration.

[36]The designation M is used when an exercise has a number of parts which have different difficulty levels.

10. *Scoring keys and rationales.* If the exercise is objectively scored (e.g., multiple-choice), please indicate the correct answer. If it is an open-end exercise or an interview or some other which requires interpretation of a response, please develop a very definite scoring key or rationale.

11. *Re-evaluation of exercise.* Finally, each exercise should be re-evaluated to see whether it meets the required criteria.

    1) If the exercise itself is reported, will the results be meaningful to both professional and lay people?

    2) Does it measure the objective for which it was intended? Is it directly related to the behavior set forth as desirable in the objective?

    3) Is it appropriate for the age level(s) indicated?

    4) Is the exercise clearly stated so that the individual will understand what he is to do? Is the vocabulary appropriate for the age?

    5) Is the exercise appropriately designated for group or individual administration?

    6) Has the estimate of difficulty level been carefully thought through in light of all the available information for the age level?

    7) Does the exercise have directionality? Is there a correct or desired response?

    8) *Is a key or scoring rationale provided as part of the exercise?*

    9) Is a time estimate included?

The development of scoring rationales and specific keys giving examples of acceptable and unacceptable responses is stressed. Following are the directions given writers:

> The need for adequate and detailed scoring rationales cannot be overstressed, particularly for open-end questions or situational tasks.
>
> A rationale states, in general terms, the intent of the exercise, in order to give the person who will be responsible for scoring an

adequate understanding of what the writer is trying to measure. In addition, a specific scoring guide which gives samples of acceptable and unacceptable responses must be prepared. In some cases specific responses can be anticipated from the "armchair." In most cases, it will be necessary to try out the exercises on actual classrooms of students (or individually with students) to get authentic responses. If this is necessary, it should be done as a part of the exercise development procedure.

The following is a sample scoring rationale of a general nature. An exercise from Citizenship asks, "Why is it a good idea for people to write letters to the President of the United States?" Acceptable responses fall into general categories such as:

1. To try to influence what he does,
2. To provide support when they approve of his action,
3. To show they care and are watching his actions.

In addition to the above types of scoring rationales, it is necessary to have specific student responses. Specific responses should be as exhaustive as possible and should anticipate marginal (or smart-alecky) responses and give directions as to how they should be treated. In most cases it will be necessary to try out the exercise in order to make this determination. Samples for a variety of exercise types follow. These samples are *not* exhaustive but serve to give some illustrations as to *type* of response and possible problems.

1. In what way are an apple and a pear alike?

   PLUS 2 points for recognition of the basic classification—fruit.

   1 point for recognition of a general classification:
   food
   to eat
   have peels
   grow
   contain vitamins

   MINUS No understanding of a classification or extraneous comments:
   same shape
   I like apples but I don't like pears.

PROBLEM RESPONSES same size (This is a classification — does it receive credit?)

2. X-rays can go through tissues of the human body but visible light cannot. Why?

PLUS Must exhibit knowledge of different wave frequencies between visible light and X-rays, or difference in wave length:

They have different wave frequencies.

X-rays go through your body because the frequency is higher and the wave length shorter than that of white light.

X-rays have a very low wave length.

MINUS Does not exhibit knowledge of way in which X-ray and visible light differ, just knowledge that they do differ:

X-rays are different from visible light.

Light cannot pass through solid objects. X-rays can go through tissue but not bone.

Visible light has small particles and particles cannot go through the body.

Here is a sample prototype exercise taken from the area of Art. The stimulus is omitted because the exercise is taken from the active pool. The remainder of the material will, however, illustrate the amount of detail which accompanies the exercise.

Directions to Administrator

If necessary clarify the second part of the exercise without indicating what aspects might give pleasure. Use a series of probes to clarify the student's responses. For example, if the response is "The colors are nice," ask "What is it about the colors that makes them nice?" If the answer is "I don't know," ask what there is about the painting which might give pleasure or pain.

Rationale

This exercise assesses (1) the degree of pleasure or pain individuals report they receive from viewing a color reproduction of a work of art, (2) whether they can describe the aspects which give rise to their pleasure or pain, (3) the range of aspects of the work which give rise to their pleasure or pain, and (4) whether there is a correspondence between judgments of pleasure or pain and their descriptions.

Directions to Scorer

The first part of the exercise is not scored. Score one point for a response in each category which corresponds to the initial judgments of pleasure or pain.

3 = three corresponding categories
2 = two corresponding categories
1 = one corresponding category
0 = no corresponding category, or no response
-1 = one non-corresponding category
-2 = two non-corresponding categories
-3 = three non-corresponding categories

Acceptable "Pleasure" Response Categories

1. Sensory, formal, and media
   The orange, yellow, and pink go nicely together.
   The shiny chrome makes me want to touch it.
   There is a beautiful repetition of shapes.
   It is nicely painted.

2. Expressive character
   The whole thing has a happy, comical look.
   It's just a fun, fun work.

3. Subject matter aspects
   The cross eyes are so funny.
   The head looks like a balloon.
   It represents a tight-lipped individual.

Acceptable "Pain" Response Categories

1. Sensory, formal, and media
   The colors are ugly together.
   The shapes don't match each other.
   All the lines are broken up.
   It is too rough.

2. Expressive character
   It's too gaudy.
   There is too much in it — too many parts — so many parts I don't even want to look at it.
   It has a messy look.

3. Subject matter aspects

The head is distorted and out of shape.

The cross eyes bother me.

If the answer is "I don't know," responses from either the pleasure or pain categories are acceptable.

Reporting

X % used three corresponding categories.
X % used two corresponding categories.
X % used one corresponding category.
X % used no corresponding categories.
X % made no response.
X % used one non-corresponding category.
X % used two non-corresponding categories.
X % used three non-corresponding categories.

The final step in Phase A is a review of the prototype exercises by the Exercise Development Advisory Group (EDAG). EDAG, as previously mentioned, is composed of measurement specialists and educators, some of whom have direct experience with inner-city and low socio-economic problems. This group reviews the prototype exercises as they are presented by the person responsible for their development (contractor representative or special consultant). They evaluate it for content validity, appropriateness, relevance and scorability. Depending upon the recommendations of this group, either further work is done to refine the prototype exercises or the production of exercises is begun.

The Preparation of Exercises

When the prototype exercises have been accepted, Phase B, or the preparation of exercises, begins.

As previously mentioned, present development of exercises is limited to creating a total of 500 minutes worth of exercises per age level. When an area is being revised, this means that the total of new materials to be written, the reserved pool (exercises already written and approved but not used in an actual assessment), and those in an assessment but not reported, should equal approximately 500 minutes.

The determination of how much new material needs to be written and which objectives need coverage depends upon: (1) which exercises are reported from the assessment and hence need to be replaced and (2) whether or not revisions in objectives have created a need for additional materials.

Once these factors have been taken into account, half the necessary new exercises are written. At least some of the independent writers who helped develop prototype exercises are asked to produce additional materials. When half the materials have been produced, including detailed directions to the administrator and detailed scoring procedures, the next review takes place. This review may be either a staff review involving outside consultants who are specialists in the area or it may be another review by EDAG. If problems are encountered they are remedied before additional exercises are written. Once this checkpoint is passed the remainder of the exercises are produced, and preparations begin for formal review of the new exercises.

Reviews

Phase C reviews consist of reviews by both subject matter specialists and lay persons. While previous reviews have been held separately for the two groups, the next Phase C reviews will consider combining the two types of reviews or at least holding them concurrently. Plans under consideration involve three basic group structures:

1. groups of subject matter specialists only
2. groups of lay persons only
3. groups containing both subject matter specialists and lay persons.

Since previous review conferences have tended to minimize interaction between specialists and lay people, it might be productive to see how mixed groups would function. One possibility would be to conduct several small review groups using all *three* types of group structures, and, following the pattern used for lay review of objectives, bring the chairmen together with contractors and special staff consultants at the end of the small group reviews.

Whatever the decision is on this, the same basic questions need to be answered: (1) Does the exercise have content validity? (2) Is it appropriate for the age group designated? (3) Is the vocabulary level appropriate? (4) Is there a danger that this exercise will offend various groups of people?

Following Phase C reviews the exercises will be further revised and refined if necessary.

Field Testing

The main purpose of field testing or tryouts is to obtain feasibility-type

information and to further develop detailed scoring keys for open-end exercises. That is, field testing is designed to answer these types of questions:

1. Can the exercise be easily understood by the assessee? Does he know what he is expected to do?
2. Is the vocabulary level such that it can be understood by low achieving students (except for reading exercises)?
3. Is the exercise biased with respect to sex or ethnic groups?
4. Are there sufficient "easy" exercises?
5. Are there problems in administration?
6. Are the time estimates reasonable?

Field testing may also be useful to test alternate ways (such as open-end vs. multiple-choice) of presenting a given exercise.

Following field testing, revisions may be made in the exercises, directions for administration or scoring keys.

Final Reviews and Selection

The process of final review and selection of exercises from the pool remains much the same as described in the previous chapter. EDAG makes a final review for the purpose of determining which exercises are eligible for selection. The regular review group is augmented by the addition of a subject matter specialist for final reviews. Exercises may be accepted as is, completely rejected or designated as needing further work. The exercises are then forwarded to the United States Office of Education where they are examined for potential invasion of privacy.

Current plans for the selection process remain unchanged, except for the addition of a subject matter person to each conference group which makes the final selection. This addition was made primarily because of recommendations made by contractors. The subject matter specialist will not be the person who is primarily responsible for the development of the exercises but will be someone who is familiar with the project and who has served as a reviewer in the Phase C reviews.

Current plans allow an eight-month period following selection for packaging and other preparations necessary for the assessment.

Summary

This chapter has presented the current scheme which is planned for the development of objectives and exercises. This plan is the result of five

years of experience and is designed to maximize the participation of subject matter specialists, educators and lay persons with a wide variety of viewpoints.

New features include the involvement of student groups in the review process, and greater emphasis on the preparation of prototype exercises and detailed scoring keys. When contracts are made with other agencies to develop areas, greater exchange between the staff and the contractor is planned, as are more checkpoints in the development of materials. Where staff directly supervises the development, the same steps as those required of the contractor will be followed.

Three years is projected as a minimum amount of time needed to develop or revise a subject area, and this assumes that no unusual problems are encountered. The development period is divided into five phases:

1. Development of objectives and prototype exercises.
2. Preparation of exercises.
3. Review and revision of exercises.
4. Field testing and revision of exercises.
5. Final reviews and selection.

GLOSSARY

AGE LEVELS [9, 13, 17, A (26-35)]. Four age groups are being assessed, each chosen to provide information at meaningful periods in the educational life of Americans. Age 9 marks the end of most students' primary education; age 13, the end of elementary education. Age 17 is usually close to the end of secondary education, and most adults have finished their formal education by the age of 26. A 10-year span of 26-35 for adults provides a large enough population from which to sample.

AIR. American Institutes for Research, a research organization with independent offices in Washington, D.C., Pittsburgh, Pennsylvania and Palo Alto, California. Currently, AIR is contracting with National Assessment to develop objectives and exercises in Citizenship and Career and Occupational Development.

ANAC. The Analysis Advisory Committee, originally called the Technical Advisory Committee (TAC). A group composed of distinguished statisticians and educational measurement specialists who advise NAEP on the most desirable and meaningful methods of processing and analyzing the data from assessment exercises.

CAPE. The Exploratory Committee on Assessing the Progress of Education (ECAPE) became the Committee on Assessing the Progress of Education (CAPE), in July 1968, when the National Assessment project began actual operation and a permanent administrative staff was formed.

CARNEGIE CORPORATION. A non-profit foundation that provided initial and continuing funds for National Assessment.

CHOICE. Each possible response in a multiple-choice exercise, including the correct answer and each distractor. (See *distractor*.)

CLUSTER. A series of exercises based on one set of stimulus material.

CODING. A coded heading for each exercise for identification purposes. It includes: (1) a contractor code, (2) the subject area, (3) an exercise number, (4) the number(s) of objective(s) being measured, (5) the age level for which it is intended, (6) estimated difficulty, (7) a code for the stimulus, (8) a code for the type of response, (9) a code indicating need for special apparatus or conditions, and (10) ages of overlap, if any.

CYCLE 1, 2, 3, ETC. In order to measure progress, a series of cycles are designed to provide comparable results for given subject areas every three

or six years. Each cycle (e.g., Cycle 1) is a six-year period during which all 10 subject matter areas are assessed either twice (Reading, Mathematics and Science) or once (all others).

DIFFICULTY LEVELS. One of three difficulty levels is assigned to each exercise, depending on the estimated percentage of people who can answer it successfully. An exercise which is considered "easy" is a 90 percent exercise (approximately 90 percent of the persons responding are expected to answer correctly), a medium difficulty exercise is called a 50 percent exercise and a "difficult" exercise is called a 10 percent exercise. M (for multiple) is used when an exercise has a number of parts which have different difficulty levels.

DIRECTIONALITY. A NAEP requirement for each exercise is that it have a "desired" response or can be scored on a scale, so that progress in a specific direction can be measured between the results of different cycles.

DISTRACTOR. An incorrect choice in a multiple-choice exercise.

ECAPE. Exploratory Committee on Assessing the Progress of Education, founded in 1964 with funds granted by the Carnegie Corporation. ECAPE was assigned to confer with teachers, administrators, school board members and interested lay people on ways to set up an assessment. (See *CAPE.*)

ECS. Education Commission of States, an organization composed of seven representatives from each of 40 states and territories. Representatives from each state include the governor, a member of each house of the legislature, and four others appointed by the governor who represent interests of state government and education. The program of ECS has five types of activities: (1) disseminating information, (2) rendering service to the states by providing specific timely information they request, (3) helping interim committees with their processes of studying a problem, (4) furnishing an opportunity for dialogue on major issues and (5) developing policies. ECS is financed by contributions from member states who are assessed according to population and per capita income. In July 1969, ECS assumed the governance of CAPE. (See *NAEP.*)

EDAG. Exercise Development Advisory Group, composed of outstanding educators, who advise the NAEP staff and contractors on the construction of exercises.

ERIE. Eastern Regional Institute for Education, located in Syracuse, New York, contracted by CAPE to collect data for the 90 Percent Study.

ETS. Educational Testing Service, a testing and research corporation in Princeton, New Jersey, which has been under contract to NAEP to develop objectives and exercises in Art, Literature, Music, Science, Social Studies and Writing.

EXERCISE. A task which is written to measure an objective. "Exercise" is used to distinguish it from the usual "test item" since an assessment exercise may be one of a variety of methods of assessment, such as performance on a musical instrument or in a discussion group, the carrying out of a scientific experiment, writing a letter or just checking a multiple-choice question.

FIELD TRYOUTS. Pretesting of NAEP exercises to obtain information regarding clarity, content validity, difficulty, timing and other factors needed in the process of review and revision of exercises. The "field" refers to any area outside the NAEP or contractors' offices where tryouts or assessment packages are administered. Field tryouts are conducted with voluntary participation from schools or individuals.

FUND FOR THE ADVANCEMENT OF EDUCATION. A non-profit foundation which has provided a substantial portion of the operating funds for the duration of the National Assessment project.

GROUP ADMINISTERED EXERCISE. Any exercise which can be administered to a group of students.

INDIVIDUALLY ADMINISTERED EXERCISE. An exercise which must be administered to one individual at a time by an interviewer.

LAY PANELS. Groups of persons who are not professional educators but are interested in education, who are asked to review National Assessment objectives and exercises.

MAG. Media Advisory Group, composed of public relations personnel and professionals in radio-television, newspapers, magazines and educational publications, who advise NAEP on the best ways to publicize its findings.

MRC. Measurement Research Center, located in Iowa City, Iowa, a research organization. MRC employs personnel and conducts the assessment in the Central and Western regions of the United States under subcontract to Research Triangle Institute. (See *RTI.*) MRC also performs the scoring and some data processing of National Assessment exercises.

NAEP. National Assessment of Educational Progress, the name assumed by CAPE when the Education Commission of the States became the governing agency of the project. (See *ECS.*)

NORC. National Opinion Research Center, a survey organization, contracted by CAPE to conduct one of the Feasibility Studies.

OBJECTIVES. A set of goals which are agreed upon as desirable directions in the education of children. NAEP objectives must be: (1) considered authentic by subject matter specialists, (2) accepted as an educational task by the school and (3) considered desirable by thoughtful lay citizens.

OPAC. Operations Advisory Committee, a group which advises NAEP on how best to implement the assessment in the field.

OPEN-END RESPONSE. An answer written by a student or interviewer in response to a stimulus. Writing in a word or phrase, writing an essay or performing a calculation are examples of open-end responses, in contrast to responses made by selecting and marking a choice in a multiple-choice question.

OUT-OF-SCHOOL 17-YEAR-OLD. Any 17-year-old person who is not enrolled in secondary school during the assessment year.

OVERLAPPING EXERCISE. An exercise which is appropriate for use at more than one age level.

PACKAGE. An assortment of National Assessment exercises for administration, typically combining two or three subject areas and several objectives within a given subject area. A package is designed so that administration time is approximately 40 minutes. Some packages are designed for groups, others for individuals. Some special packages contain scientific apparatus or instruments for the student to illustrate his skills.

PC. The Psychological Corporation, a New York-based testing and research organization responsible for the initial development of mathematics objectives and exercises.

RTI. Research Triangle Institute, a research organization located in Research Triangle Park in North Carolina. RTI is responsible for sampling and for the conducting of the assessment across the country. RTI employs exercise administrators, district supervisors and regional supervisors for the Northeastern and Southeastern regions of the country and subcontracts with Measurement Research Center (MRC) for the administration in the Central and Western regions. (See *MRC*.)

SEL. Southeastern Education Laboratory in Atlanta, Georgia, contracted by NAEP to collect data for the 90 Percent Study.

SES. Socio-educational status. In National Assessment terminology, an index that attempts to classify individuals on the basis of being privileged or under-privileged, from the point of view of both sociological and educational opportunities available to an individual.

SHELVE. To withdraw an exercise from the active pool of exercises. The exercise may receive reconsideration at a later date.

SRA. Science Research Associates, a subsidiary of IBM, originally contracted to develop the objectives and exercises in Reading.

STEM. That portion of an exercise which states the problem or asks the question.

SUBJECT MATTER AREAS. Areas selected for national assessment are Art, Career and Occupational Development (formerly called Vocational Education), Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies and Writing (written expression). Additional areas are to be developed in future years.

TAC. Technical Advisory Committee. (See *ANAC*.)

USOE. United States Office of Education, a division of the Department of Health, Education and Welfare.

YEAR 01, 02, 03, ETC. A sequential number is assigned to each "year" of assessment activities in the field. Year 01 was March 1969 to February 1970. All other assessment periods (Year 02, 03, etc.) will be from October to August. Thus, Year 02 is from October 1970 to August 1971.

REFERENCES

Berdie, Frances S. To rile your community, ask questions like these. *American School Board Journal*, June,1970.

Department of Elementary School Principals, NEA. *National assessment of educational progress: some questions and comments*. (Rev. ed.) Washington, D.C.: Author, 1968.

Gordon, George B. *National assessment feasibility study*. Unpublished report to Exploratory Committee on Assessing the Progress of Education from Educational Testing Service, June, 1967.

Johnstone, John W. C. and Spaeth, Joe L. *Administration of test exercises in the home*. Unpublished report to Exploratory Committee on Assessing the Progress of Education from National Opinion Research Center, September, 1967.

Knapp, Thomas R. *The choices study*. Unpublished report, Exploratory Committee on Assessing the Progress of Education, 1968.

Knapp, Thomas R. *The mathematics study*. Unpublished report, Exploratory Committee on Assessing the Progress of Education, 1968.

Mager, Robert F. *Preparing instructional objectives*. Palo Alto, California: Fearon Publishers, 1962.

Merwin, Jack C. and Womer, Frank B. Evaluation in assessing the progress of education to provide bases of public understanding and public policy. In *NSSE yearbook. Education evaluation: new roles, new means*. Chicago, Illinois: University of Chicago Press, 1969.

Tyler, Ralph W. Introduction. In E. L. Norris (Ed.), *National assessment of educational progress: science objectives*. Ann Arbor, Michigan: Committee on Assessing the Progress of Education, 1969.

Tyler, Ralph W. Progress report on the try-outs of the assessment exercises. Unpublished paper. St. Paul, Minn.: Exploratory Committee on Assessing the Progress of Education, July, 1967.

Womer, Frank B. Research toward national assessment. *Western Regional Conference on Testing Problems, Proceedings*, 1968, 34-49.

Womer, Frank B. *What is national assessment?* Ann Arbor, Michigan: National Assessment of Educational Progress, 1970.